Notas de ECONOMETRÍA

Doble Grado en Ingeniería Informática y ADE,

ETSIINF, UPM

Juan J. Morales-Ruiz

6 de mayo de 2019

II

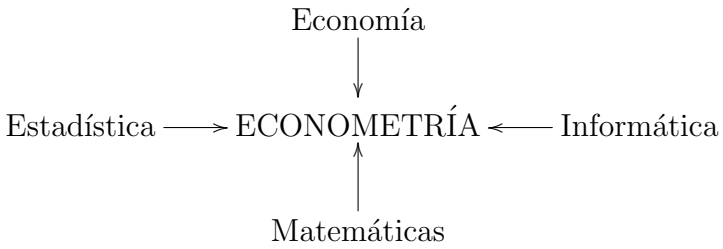
Índice general

Capítulo 1

Introducción

1.1. ¿Qué es la econometría?

La econometría (economía+medida) tiene por objetivo utilizar métodos matemáticos y estadísticos para el estudio *cuantitativo* de los modelos económicos con la finalidad de evaluarlos, hacer predicciones y tomar decisiones. Para ello plantea un *modelo econométrico* a partir de un modelo económico. Además, para el tratamiento de los datos usualmente la econometría utiliza herramientas informáticas.



Ejemplos del tipo de problemas que se tratan en econometría: estudiar las causas de las crisis y tratar de predecirlas o minimizar sus efectos, evolución de los mercados de valores, del consumo respecto a la renta, dependencia del precio de las viviendas respecto a su superficie, etc. También podrían estudiarse fenómenos no económicos, por ejemplo, la evolución anual del nivel de los pantanos, la evolución de la temperatura del planeta y si tiene relación con la actividad humana (p. ej., emisión de gases) o con otros factores, como los ciclos de la actividad solar, etc.

Un *modelo económico* es un modelo que plantea la dependencia de una (o varias) variable(s) dependiente(s) y, en función de otras variables independientes (o *variables explicativas*) $x_1, ..., x_k$,

$$y = f(x_1, ..., x_k).$$

El modelo es matemático *determinista*, en el sentido de que las variables no son aleatorias. A la variable y se le llama también *variable explicada* y a las $x_1, ..., x_k$ *variables explicativas*, ya que si el modelo está correctamente formulado, las variables independientes explican adecuadamente el comportamiento de la variable y. El modelo puede precisar o no la forma exacta de la función f (por ejemplo que sea lineal) pero, en cualquier caso, nos tendría que decir algunas de sus propiedades (por ejemplo, que sea creciente respecto a algunas de las variables explicativas).

Un *modelo econométrico* es un modelo que se construye a partir de un modelo económico (o mediante otros mecanismos, como una muestra de datos), pero con dos diferencias respecto al modelo económico: se precisa la forma de la función f y, en general, las variables son aleatorias, es decir el modelo no es determinista sino estadístico.

Ejemplo. Según Keynes el consumo y de una familia de una población está relacionado con la renta x de forma lineal

$$y = \alpha + \beta x, \tag{1.1}$$

con α y β números reales. Este es un modelo económico. Un posible modelo econométrico correspondiente sería

$$y = \alpha + \beta x + \varepsilon, \tag{1.2}$$

con x, y y ϵ, variables aleatorias (aunque también x se podría considerar una variable matemática, ie, no aleatoria). Está clara la ventaja que tiene este modelo sobre el modelo determinista (1.1): la variable ϵ da cuenta de otras posibles variables en el consumo, que difícilmente pueden cuantificarse de manera precisa o que complicarían enormente el modelo; es decir, obviamente ante una misma renta dos unidades familares no consumen lo mismo. Naturalmente, el modelo econométrico también tiene que precisar qué tipo de variable aleatoria es la variable ε.

Fases en la construcción de un modelo econométrico:

1) *Planteamiento (o especificación) del modelo.* A partir de un modelo económico y de una muestra de datos de las variables se plantea el modelo econométrico.

2) *Estimación de los parámetros del modelo.* A partir de una muestra de datos y de métodos estadísticos, se estiman los parámetros del modelo.

3) *Validación (o diagnosis) del modelo.* Se estudia si el modelo representa adecuadamente los fenómenos económicos que se tratan de estudiar. En caso de no pasar la diagnosis habría que modificar el modelo.

Ejemplo. En el ejemplo del consumo de Keynes la ecuación (1.2) da la especificación del modelo, afirmando que, por ejemplo,

$$\varepsilon \sim N(0, \sigma^2).$$

El modelo tiene 3 parámetros α, β y σ, que se estimarían a partir de una muestra de datos.

Tipos de datos:

1) *Datos de corte transversal*: datos que no están ordenados cronológicamente, o donde el orden temporal no es relevante; esto es así si, en particular, se han obtenido durante (aproximadamente) un mismo instante de tiempo. Por ejemplo, el consumo de un conjunto de familias durante un mes.

2) *Datos de series de tiempo*: datos obtenidos cronológicamente en el tiempo y donde el tiempo es relevante. Por ejemplo, la evolución del precio de las acciones lo largo de un año o la del PIB de un país a lo largo de varios años.

3) *Datos de panel*: se mezclan datos de corte transversal con datos de series de tiempo. Por ejemplo, los salarios de una serie de individuos, respecto a sus años de educación y su antiguedad en el puesto de trabajo, a lo largo de varios años. En este curso no trataremos con datos de panel.

1.2. Prerrequisitos

De matemáticas necesitaremos: álgebra lineal (más concretamente, álgebra de matrices) y cálculo diferencial e integral de varias variables.

De estadística necesitaremos: probabilidad condicionada, independencia, vectores aleatorios (ie, distribuciones de variables aleatorias multivariantes), en particular, matriz de covarianzas de un vector aleatorio, la distribución normal simple y multivariante, distribuciones asociadas a la normal (χ^2, t de Student, F de Fisher-Snedecor), muestras, teorema central del límite, estimación, intervalos de confianza y contrastes de hipótesis. En un cierto sentido la econometría puede considerarse como un laboratorio de la estadística, en donde aplicamos prácticamente todos los conceptos básicos de estadística. El valor esperado de una variable aleatoria x lo designamos también como su media o su esperanza $E[x]$. Vamos a recordar brevemente algunos de estos conceptos.

1.3. Distribuciones multivariadas

Un vector aleatorio (o variable aleatoria multivariante de dimensión n) es un vector $(x_1, ..., x_n)$ de variables aleatorias definidas sobre el mismo espacio muestral. A partir de ahora, si no se dice lo contrario, supondremos que las variables aleatorias son continuas, entonces el vector aleatorio tiene asociada una función densidad $f(x_1, ..., x_n)$, que permite calcular la probabilidad sobre $A \subset \mathbb{R}^n$:

$$P(A) = \int_A f(x_1, ..., x_n)\, dx_1 \cdots dx_n.$$

Las densidades marginales, $f(x_i)$ se obtienen integrando respecto a todas las otras variables. Las densidades marginales definen las distribuciones de probabilidades marginales de las variables x_i. Las variables $x_1, ..., x_n$ son independientes $\Leftrightarrow f(x_1, ..., x_n) = f(x_1) \cdots f(x_n)$.

Def. Una matriz simétrica cuadrada $A = (a_{ij})$, i, $j = 1, ..., k$ es *definida positiva* si para todo vector no nulo, $\mathbf{v} \in \mathbb{R}^k$, se cumple $\mathbf{v}'A\mathbf{v} > 0$ y se llama *semidefinida positiva* si $\mathbf{v}'A\mathbf{v} \geq 0$.

Se puede demostrar que A (simétrica) es definida positiva si los k menores

$$a_{11}, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}, ..., |A|$$

son todos positivos (criterio práctico). Esto equivale a que todos sus autovalores sean positivos (¿por qué?). Análogamente, se obtiene un criterio de que A sea semidefinida positiva, permitiendo la igualdad en las afirmaciones anteriores.

Recordemos que la covarianza $\sigma_{ij} = \text{cov}(x_i, x_j) = E[(x_i - E[x_i])(x_j - E[x_j])] = E[x_i\, x_j] - E[x_i]E[x_j]$, siendo $\sigma_{ii} = \sigma_i^2$ la varianza de x_i. Si x_i, x_j son independientes, entonces $\sigma_{ij} = 0$. La *matriz de covarianzas* (o de varianzas-covarianzas) del vector aleatorio $\mathbf{x} = (x_1, ..., x_n)$ se define como

$$\text{var}(\mathbf{x}) = \Sigma = (\sigma_{ij}),$$

que es una matriz simétrica y semidefinida positiva. Si $x_1, ..., x_n$ son independientes $\Rightarrow$ la matriz Σ es diagonal.

El *coeficiente de correlación* entre x_i y x_j viene definido por

$$\rho_{ij} = \frac{\text{cov}(x_i, x_j)}{\sigma_i \sigma_j},$$

con $\sigma_i^2 = V[x_i]$ (varianza de x_i, que también se escribe $\text{var}(x_i)$). Dos variables aleatorias están *incorreladas* si su coeficiente de correlación es cero, ie, lo es su covarianza.

La *distribución normal multivariante* $N(\boldsymbol{\mu}, \Sigma)$ con vector de medias

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$$

y matriz Σ, simétrica y definida positiva, viene definida por un vector aleatorio, que ahora escribimos como vector columna

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

de función densidad

$$f(x_1, ..., x_n) = \frac{1}{(\sqrt{2\pi})^n \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right), \qquad (1.3)$$

siendo $(\mathbf{x}-\boldsymbol{\mu})'$ el vector fila transpuesto del vector columna $\mathbf{x}-\boldsymbol{\mu}$. Entonces sus esperanzas marginales serán μ_i y su matriz de covarianzas será Σ. Fijémonos que estamos suponiendo que la matriz de covarianzas es regular, eso es así siempre, ya que hemos supuesto que es definida positiva. Además, se cumple que las distribuciones marginales son también normales con las medias y desviaciones típicas las esperadas

$$x_i \sim \mathrm{N}(\mu_i, \sigma_i).$$

Observación. En estadística consideraremos los vectores siempre como vectores columna, pues trabajaremos con las aplicaciones lineales como matrices. Dada una matriz X, a la transpuesta se denota en estadística como X' (¡no confundir con la derivada!). También es frecuente en econometría escribir las variables aleatorias en minúsculas (debido a que las mayúsculas las reservamos para denotar matrices); el contexto hace que no se confunda con una variable matemática, por ej., si escribimos

$$\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma),$$

claramente nos referimos al vector aleatorio que tiene una distribución normal multivariante, por el contrario el vector $\mathbf{x}$ que aparece en su densidad (1.3) es un vector de variables matemáticas.

Si $\mathbf{x}$ es un vector aleatorio normal, entonces cualquier vector aleatorio que dependa linealmente de este también es normal, ie,

$$\mathbf{y} = \mathbf{y}_0 + A\mathbf{x},\ \mathbf{y}_0 \in \mathbb{R}^n,\ A \ \text{ matriz } m \times n \text{ de números reales}, \qquad (1.4)$$

es un vector aleatorio normal (¿por qué?)). Luego para identificarlo sólo hemos de conocer su esperanza y su matriz de covarianzas.

Otra propiedad de la distribución normal multivariante es que la independencia se puede obtener a partir de la matriz de covarianzas, ie, si el vector aleatorio tiene una distribución normal multivariante, entonces: las variables x_i y x_j son independientes, $i, j = 1, ..., n \Leftrightarrow$ la matriz de covarianzas es diagonal.

Ejercicio I.1. Demostrar la propiedad anterior.
Solución. Sólo hemos de demostrar $\Leftarrow$, ya que $\Rightarrow$ es conocido. Si Σ es diagonal, la función densidad (1.3) será

$$f(x_1,...,x_n) = \frac{1}{(\sqrt{2\pi})^n \sigma_1 \cdots \sigma_n} \exp\left(-\frac{1}{2}\left[\sum_{i=1}^{n}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right]\right) = f(x_1)\cdots f(x_n).$$

A partir de ahora una variable aleatoria la designaremos frecuentemente por v.a. (o incluso por va).

Finalmente recordamos las tres distribuciones relacionadas con la normal, que se usan frecuentemente en inferencia. Una *muestra aleatoria simple* de una población x (v.a.), con desviaciòn típica σ, viene definida por un vector aleatorio $\mathbf{x}$ con matriz de covarianzas $\Sigma = \sigma^2 I$ (I matriz identidad), debido a que, por definición, los elementos del vector aleatorio cumplen $x_i \sim x$ y son independientes dos a dos. En ocasiones una muestra aleatoria simple de una distribución normal se denota como

$$x_i \sim \text{ NID}(\mu, \sigma),\ i = 1, ..., n \tag{1.5}$$

(NID="Normal Independientemente Distribuidas").

En las definiciones que vienen a continuación la población de las muestras será siempre $x \sim \text{N}(0,1)$ (distribución normal estándar). Se definen

1) *Distribución* χ^2_n. Dada la muestra simple

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

se define

$$\chi^2_n \sim \sum_{i=1}^{n} x_i^2,$$

que también se escribe como $\chi^2(n)$. Su valor esperado es n y su varianza $2n$. Por tanto, aumenta la media y la dispersión con n.

2) *Distribución* t_n. Dadas $x \sim \text{N}(0,1)$, $y \sim \chi^2_n$ independientes, se define

$$t_n \sim \frac{x}{\sqrt{\frac{y}{n}}},$$

que también se escribe como $t(n)$. Para $n > 2$, su media (valor esperado) es 0 y su varianza $\frac{n}{n-2}$. Si $n \to \infty$, t_n tiende a la normal estándar (ya que su función densidad tiende a la de esta última). Además, su densidad es una función par.

3) *Distribución* F_{n_1,n_2}. Sean $y_1 \sim \chi^2_{n_1}$, $y_2 \sim \chi^2_{n_2}$ independientes, se define

$$F_{n_1,n_2} \sim \frac{y_1/n_1}{y_2/n_2},$$

que también se escribe como $F(n_1, n_2)$. Observamos que la distribución $F_{1,n}$ coincide con la distribución t^2_n.

Ejercicio I.2.

a) Obtener el valor esperado de la distribución χ^2_n.

b) Sabiendo que $E[x^4] = 3$, para x distribución normal estándar, obtener la varianza de χ^2_n.

Solución.

a) Si $x \sim \mathrm{N}(0, 1)$, $E[x^2] = V[x] = 1 \Rightarrow E[\chi^2_n] = n$, al ser una suma de va independientes.

b) $V[\chi^2_n] = \sum_{i=1}^n V[x_i^2] = n(E[x^4] - E[x^2]) = 3n - n = 2n$.

El siguiente ejercicio está sacado de [1] (ejercicio 1.5,a)

Ejercicio I.3. Sea $\mathbf{x}$ un vector aleatorio de dimensión n definido por una muestra de una población $x \sim \mathrm{N}(0, 1)$. Demostrar que la variable aleatoria $z = \mathbf{x}'P\mathbf{x}$, con P una matriz cuadrada simétrica $n \times n$ idempotente, $P^2 = P$, de rango r sigue una distribución $z \sim \chi^2_r$ (este tipo de matrices P son lo que llamaremos proyectores ortogonales).

Solución. (La solución que damos requiere conocer diagonalización ortogonal) Primero, es fácil demostrar que los autovalores de P son 1 (con multiplicidad r) y 0 (con multiplicidad $n - r$). Además, la matriz P diagonaliza en una base ortogonal $\Rightarrow z = \mathbf{y}'D\mathbf{y}$, con $\mathbf{y} = R'\mathbf{x}$, R matriz ortogonal (de hecho es una rotación) $R^{-1} = R'$ y, ordenando los autovalores si es necesario, D es una matriz con 1 en los primeros r elementos de la diagonal, siendo ceros el resto de sus elementos de matriz. Por tanto, $z = \sum_{i=1}^r y_i^2$. Basta

demostrar que cada y_i sigue una distribución normal estándar. Teniendo en cuenta que y_i es una combinación lineal de las distribuciones x_i con coeficientes definidos por un vector unitario (fila i-ésima de R'), sabemos que es normal, siendo tanto su valor esperado como su varianza los adecuados, usando la independencia de las x_i.

Capítulo 2

Regresión lineal simple

2.1. Modelo matemático

2.1.1. Mínimos cuadrados

Problema. *Dado un conjunto de puntos en el plano* $\mathbb{R}^2$, (x_i, y_i), $i = 1, ..., n$ *(muestra o "nube de puntos") , obtener la recta*

$$y = a + bx \tag{2.1}$$

que aproxima mejor ese conjunto de puntos.

A la variable x se le llama variable *explicativa* (independiente, regresora o exógena), a la y, variable *explicada* (a ser explicada, dependiente o endógena). Los *residuos* (o residuales) serán

$$e_i = y_i - (a + bx_i).$$

Para resolver el problema anterior utilizamos el *método de mínimos cuadrados* (o método ordinario de mínimos cuadrados): minimizar la suma de los residuos al cuadrado

$$\sum_{i=1}^{n} e_i^2.$$

Es decir, calcular el mínimo de la función de dos variables

$$g(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

Observamos que el problema no es simétrico en las variables x, y, ie, resolviendo x respecto a y en (2.1) obtenemos una recta cuyos errores al cuadrado (ahora en x) serían distintos a los de la variable y. A veces en un problema concreto hay que decidir qué variable tomamos como explicadora.

Obtengamos dicho mínimo,

$$\begin{aligned} \frac{\partial g}{\partial a} &= -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \\ \frac{\partial g}{\partial b} &= -2\sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0, \end{aligned} \tag{2.2}$$

es decir,

$$\begin{aligned} an + b\sum x_i &= \sum y_i \\ a\sum x_i + b\sum x_i^2 &= \sum x_i y_i \end{aligned} \tag{2.3}$$

(estas ecuaciones se llaman *ecuaciones normales*), de solución

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} n\bar{y} \\ \sum x_i y_i \end{pmatrix}, \tag{2.4}$$

$$\begin{aligned} a &= \frac{\bar{y}\sum x_i^2 - \bar{x}\sum x_i y_i}{\sum x_i^2 - n\bar{x}^2} \\ b &= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \end{aligned} \tag{2.5}$$

siendo $\bar{x}$ e $\bar{y}$ las medias muestrales de x e y. Se observa que b es el cociente de la covarianza muestral de x e y, dividida por la varianza muestral de x. Aquí hemos llamado varianza muestral a $1/n\sum(x_i - \bar{x})^2$, no hay acuerdo respecto a la terminología: en algunas referencias la varianza muestral se define como $\frac{1}{n-1}\sum(x_i - \bar{x})^2$ (debido a que este último es un estimador insesgado de la varianza poblacional).

De la primera ecuación de (2.2) obtenemos otra expresión para a

$$a = \bar{y} - b\bar{x},$$

que facilita el cálculo, pues obtenemos a de b. Además, indica que *la recta buscada pasa por el punto* $(\bar{x}, \bar{y})$.

En el desarrollo anterior se ha supuesto que el sistema lineal (3.7) es compatible, lo que equivale a que haya variación muestral en x, es decir, a que todos los valores x_i no sean el mismo, en efecto,

$$\begin{vmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{vmatrix} = n(\sum x_i^2 - n\bar{x}^2) = n\sum (x_i - \bar{x})^2 = \Delta \tag{2.6}$$

(es decir, un múltiplo de la varianza muestral de x). Por tanto Δ siempre será positivo, salvo que la varianza muestral de x sea nula (¿qué pasa gráficamente si $n > 1$ pero x no tiene variación muestral?).

Verifiquemos que es un mínimo. La matriz Hessiana es

$$\begin{pmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum x_i^2 \end{pmatrix};$$

que es definida positiva si $n > 0$ y su determinante 4Δ, con Δ el determinante (2.6), que será mayor que cero si existe variación muestral en la variable explicativa x. Por tanto, bajo la hipótesis de que x tenga variación muestral, el punto obtenido en (2.5) es un mínimo.

A la recta obtenida se le llama *recta de regresión* de la muestra de datos (x_i, y_i). Como la recta de regresión pasa por el punto $(\bar{x}, \bar{y})$, también la podemos escribir como

$$y - \bar{y} = b(x - \bar{x}).$$

Al parámetro a también se le llama *intercepto.*

2.1.2. Bondad del ajuste

Ejercicio II.1. Demostrar que los residuos cumplen

$$\sum e_i = 0, \sum (x_i - \bar{x})e_i = 0.$$

Solución. La primera ecuación de (2.2) nos dice que $\sum\ e_i = 0 \Rightarrow \sum \bar{x}e_i = 0$, la segunda que $\sum x_i e_i = 0$.

Los cuadrados $\sum e_i^2$ dan una medida de cuan mala es la aproximación de la nube de puntos por la recta. Se cumple

$$y_i - \bar{y} = b(x_i - \bar{x}) + e_i,$$

$$\sum (y_i - \bar{y})^2 = b^2 \sum (x_i - \bar{x})^2 + \sum e_i^2, \tag{2.7}$$

debido al último ejercicio. La igualdad (2.7) se denota como $SST = SSE + SSR$, con

$$\begin{aligned} SST &= \textit{la suma total de cuadrados} \text{ (“total sum of squares”)} \\ SSE &= \textit{la suma explicada de cuadrados}\text{(“explained sum of squares”)} \\ SSR &= \textit{la suma residual de cuadrados}\text{(“ sum of squares residuals”),} \end{aligned}$$

que se interpreta como que las desviaciones respecto a la media en las ordenadas y, SST, se compone de una parte explicada a través de las desviaciones en la variable explicativa x y una parte “no explicada”, la de los residuos.

Si la parte explicada es grande, esto indica un buen ajuste y bajos valores de los residuos. Entonces una medida *adimensional* de la bondad del ajuste viene dada por el *coeficiente de determinación*

$$R^2 = \frac{SSE}{SST} = \frac{b^2 \sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}, \tag{2.8}$$

que por (2.5), también se puede escribir

$$R^2 = \frac{(\sum (x_i - \bar{x})(y_i - \bar{y}))^2}{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}.$$

Se observa que $0 \leq R^2 \leq 1$ y que lo que hace el método de mínimos cuadrados es obtener el valor máximo de R^2 (ie, mínimo de SSR). Además, (2.8) nos dice que R^2 nos da *el porcentaje de la variación muestral de y que es explicada por x*: $100R^2$ será ese porcentaje.

Si R^2 es cercano a 1 el ajuste es bueno, y si es cercano a 0, malo. Fijémonos que b cercano a cero contribuye a un mal ajuste, es decir rectas con poca pendiente darán, en general, mal ajuste.

El coeficiente de correlación lineal de Pearson es la raíz cuadrada R del coeficiente de determinación, pero hay una ambiguedad en el signo: R tiene el signo de la pendiente de la recta de regresión.

La utilidad de la regresión está en su capacidad de predicción: $y = a + bx$ sería el valor que “predice”la regresión para la variable explicada, para un valor x de la variable explicativa.

Ejercicio II.2. Una empresa desea conocer cómo depende el coste de un nuevo modelo de viga respecto su resistencia. Para ello realiza un experimento con cuatro vigas de diferente coste, siendo los resultados los de la tabla adjunta.

Resistencia (Tm)	Precio (miles de euros)
1.2	2.4
2	4
3.5	4.5
7.4	9

Obtener la recta de regresión de los siguientes datos con $n = 4$, estudiando además la bondad del ajuste.

Solución. Para evitar errores de redondeo, vamos a tomar 4 cifras decimales significativas. Las medias muestrales son $\bar{x} = 3.525$, $\bar{y} = 4.975$. Entonces

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = 1.0147,\ a = \bar{y} - b\bar{x} = 1.3981.$$

La recta de regresión es

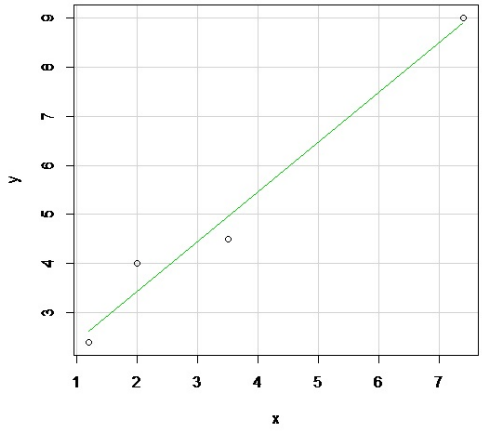
$$y = 1.3981 + 1.0147x.$$

Además

$$SSE = b^2 \sum(x_i - \bar{x})^2 = 23.4224,\ SST = 24.0075,\ R^2 = \frac{SSE}{SST} = 0,9756,$$

es decir, la resistencia de la viga explica el 97.56 por ciento de la variación muestral de su precio.

El ajuste es muy bueno, como también era de esperar gráficamente.



Una crítica al resultado de este ejercicio es que, aunque el ajuste es muy bueno, la nuestra tiene un tamaño pequeño $n = 4$, que en el caso de las vigas podía estar justificado, pues se ha de destruir la viga para realizar el experimento. Es intuitivo que el temaño de la muestra debería incidir en la fiabilidad del estudio. El caso extremo es para $n = 2$, con variación (ie, $x_1 \neq x_2$), obteniendo siempre $R^2 = 1$ (¿por qué?).

2.1.3. Regresiones no lineales que se reducen a lineales

Si la relación entre la variable y y x es tal que con un cambio de variable la convertimos en lineal, es posible también aplicar el método precedente. Usualmente, se opta por realizar un determinado cambio de variable observando el comportamiento de la nube de puntos. Dos ejemplos:

1) *Regresión logarítmica.* $y = a+b\log x$. Aplicamos el metodo a la muestra $(logx_i, y_i)$.

2) *Regresión exponencial.* $y = ce^{bx}$, $c > 0 \Leftrightarrow \log y = \log c + bx$. El método se aplica a $(x_i, \log y_i)$. Obtenemos c como $c = e^a$.

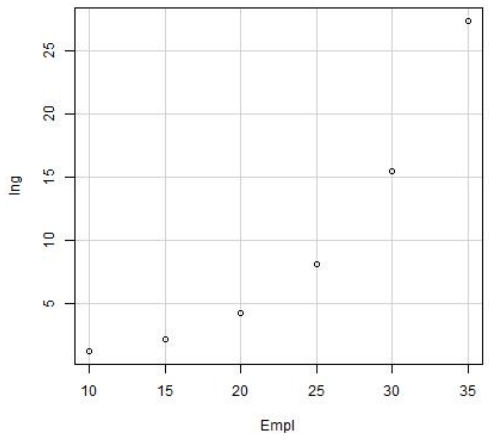
Es fácil obtener más casos, p. ej., $y = d + cx^b$, $c > 0$ (¡ejercicio!).

Ejercicio II.3. Una empresa desea conocer la dependencia de sus ingresos netos en función del número de sus empleados y posee los siguientes datos de los últimos 6 años, donde los ingresos se miden en cientos de miles de euros.

Empleados	Ingreso
10	1.22
15	2.13
20	4.26
25	8.12
30	15.40
35	27.35

Obtener una curva de regresión adecuada y estudiar la bondad del ajuste. Comparar el resultado con la bondad del ajuste para una regresión lineal realizada directamente a partir la tabla anterior (es decir, sin cambio de variable previo).

Idea Sol.: La gráfica de dispersión es



Parece claro lo que habría que hacer...

2.2. Modelo estadístico

Como en otras áreas de aplicación de la estadística, esencialmente, los únicos datos de que dispone un econometrista para realizar sus estudios son muestras. En el caso de la regresión simple estos datos son parejas de datos numéricos que representan dos variables de una cierta "población". Si tomamos otra muestra, los resultados no serán los mismos, por tanto, es razonable considerar las variables como aleatorias. Una vez que el modelo ha sido planteado (o especificado), dos partes esenciales del estudio son la estimación e inferencia.

En el modelo matemático (no aleatorio) de la regresión hemos considerado solo la aproximación de los datos por mínimos cuadrados, pero obviamente si realizamos otro experimento muestral el resultado no va a ser el mismo, y por tanto, la curva de regresión tampoco será exactamente la misma. Es decir, debemos pasar a un modelo estadístico, considerando las propiedades de la población a estudiar como variables aleatorias. Una forma de modelar esto en el caso de una recta de regresión es pensar en una ecuación del tipo

$$y = \alpha + \beta x + \varepsilon \text{ ("modelo poblacional")}, \tag{2.9}$$

siendo α y β parámetros (es decir, números reales, no v.a.) y x, y y ε v.a.. Según este modelo la perturbación aleatoria "oculta" ε, que será una suma

de fenómenos variados incontrolados, es la responsable de las variaciones observadas en las muestras. Las hipótesis que se hacen tratan de precisar esta idea.

2.2.1. Especificación

El modelo se basa en las siguientes **Hipótesis**:

(H1) *Linealidad.* Los datos se generan mediante una muestra

$$y_i = \alpha + \beta x_i + \varepsilon_i,\ i = 1, ..., n,$$

siendo α y β constantes desconocidas (intercepto y pendiente) y x_i constantes conocidas que, por tanto, tendrán los mismos valores en cualquier otra muestra. Como en cualquier teoría estadística de muestras, el tamaño n también está fijado.

(H2) *Variación muestral de la variable explicativa.* No todos los valores de x_i son iguales.

(H3) *Perturbación con media cero.* La perturbación (o error) ε_i, $i = 1, ..., n$ es un vector aleatorio con $E[\varepsilon_i] = 0$.

(H4) *Homocedasticidad.* La varianza de la perturbación es constante en la muestra $V[\varepsilon_i] = \sigma^2$, $i = 1, ..., n$.

(H5) *No correlación.* Los pares de perturbaciones están incorrelados dos a dos $E[\varepsilon_i \varepsilon_j] = 0$, $i \neq j$, $i, j = 1, ..., n$.

A este modelo se le llama *modelo de Gauss-Markov* de la regresión (simple). Usualmente se supone también que la perturbación es una distribución normal

(H6)

$$\varepsilon_i \sim \text{ NID}(0, \sigma).$$

Con las hipótesis anteriores se cumple

$$E[y_i] = \alpha + \beta x_i,\ V[y_i] = \sigma^2,\ \text{cov}(y_i, y_j) = 0\ (i \neq j) \Rightarrow \text{var}(\boldsymbol{\varepsilon}) = \text{ var}(\mathbf{y}) = \sigma^2 I,$$

siendo var($\boldsymbol{\varepsilon}$), (var($\mathbf{y}$)), las matrices de covarianzas de los vectores aleatorios

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

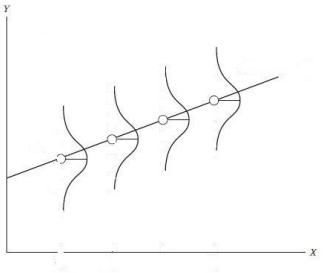
respectivamente.

Teniendo en cuenta el modelo poblacional (2.9), (H3) no es realmente restrictiva (ie, lo podemos suponer siempre): basta ajustar el parámetro de intercepto α a

$$\alpha + E[\varepsilon] = E[y] - \beta x.$$

Además, como y_i es normal,

$$y_i \sim \text{ NID}(\alpha + \beta x_i, \sigma)$$



Ejercicio II.4. Demostrar con detalle las afirmaciones anteriores.

En lugar de suponer que las variables x_i no son aleatorias, algunas referencias suponen que son v.a., pero que tratamos con probabilidades condicionadas en las distribuciones de ε_i y y_i, respecto a valores fijos (ie, no aleatorios) del vector aleatorio

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Por ejemplo, la hipótesis (H4) sería $E[\varepsilon_i | x_i] = 0$ (abuso de notación: ponemos x_i aquí para indicar un valor fijo de la v.a. x_i).

2.2.2. Estimación

Por estimación entenderemos siempre estimación puntual. Recordemos que un *estimador* de un parámetro θ de una v.a. x es un estadístico $\hat{\theta} = T(x_1, x_2, ..., x_n)$ obtenido a partir de una muestra de x, que permite realizar una *estimación* del parámetro al sustituir los valores empíricos de una muestra concreta. Luego un estimador es una v.a. y una estimación es una constante (un número real), precisamente debido a que esta constante cambia de muestra a muestra es lo que hace que un estimador sea una v.a.. Aquí seguiremos la tradición estadística de *no distinguir en la notación un estimador de su estimación, denotando ambos como* $\hat{\theta}$; la distinción se hace por el contexto.

El modelo anterior tiene tres parámetros α, β y σ. La estimación de los mismos se puede hacer de varias formas, aquí vamos a hacerlo por el método de mínimos cuadrados; es decir, dada la muestra (x_i, y_i), con x_i constantes fijas conocidas e y_i v.a., cumpliendo las propiedades (H1)-(H6), estimamos α y β mediante las fórmulas (2.5)

$$\begin{aligned} \hat{\alpha} &= \frac{\bar{y}\sum x_i^2 - \bar{x}\sum x_i y_i}{\sum x_i^2 - n\bar{x}^2} \\ \hat{\beta} &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}. \end{aligned} \tag{2.10}$$

Estos estimadores son v.a., al depender de la muestra y_i.

Los estimadores anteriores se pueden expresar mediante combinaciones lineales de las perturbaciones:

Ejercicio II.5

a) Demostrar que

$$\hat{\beta} = \beta + \sum c_i \varepsilon_i, \text{ con } \quad c_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}; \tag{2.11}$$

$$\hat{\alpha} = \alpha + \sum d_i \varepsilon_i, \text{ con } \quad d_i = \frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}. \tag{2.12}$$

b) Demostrar que

$$\sum c_i = 0, \quad \sum c_i^2 = \frac{1}{\sum (x_i - \bar{x})^2}; \tag{2.13}$$

$$\sum d_i = 1, \quad \sum d_i^2 = \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}. \tag{2.14}$$

Solución. Se basa en una aplicación reiterada de la identidad $\sum (x_i - \bar{x}) = 0$.

a) Por (2.10) y (H1)

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})(\alpha + \beta x_i + \varepsilon_i)}{\sum (x_i - \bar{x})^2} =$$

$$\beta + \sum c_i \varepsilon_i.$$

Mediante la expresión anterior,

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \alpha + \beta\bar{x} + \bar{\varepsilon} - \hat{\beta}\bar{x} = \alpha + (\beta - \hat{\beta})\bar{x} + \frac{1}{n}\sum \varepsilon_i =$$

$$\alpha - \bar{x}\sum c_i \varepsilon_i + \frac{1}{n}\sum \varepsilon_i = \alpha + \sum d_i \varepsilon_i.$$

b) Se obtiene a partir de las expresiones de c_i y d_i dadas en a).

EJERCICIO CON ORDENADOR II.A: ejercicios II.2 y II.3. Utilizando ahora un programa informático (R y el paquete R Commander), estimar el intersecto, la pendiente, la bondad del ajuste y la gráfica de la recta de regresión, junto con la nube de puntos, para los datos dados en los ejercicios II.2 y II.3 de la sección 2.1.

EJERCICIO CON ORDENADOR II.B: simulación de un modelo.

a) Generar mediante un programa informático el siguiente modelo estadístico de regresión lineal simple

$$y_i = 3 + 2x_i + \varepsilon_i, \quad \varepsilon_i = \text{NID}(0, 1), \quad i = 1...100,$$

siendo $\mathbf{x}' = (x_1, ..., x_{100}) = (1, ..., 1, 2..., 2, ..., 10, ...10)$, apareciendo 10 veces cada uno de los dígitos del 1 al 10.

b) Estudiar el ajuste del modelo, obteniendo la estimación del intercepto y la pendiente, junto con la gráfica de la nube de puntos y la recta de regresión.

c) Aumentar el tamaño de la muestra a $n = 500$, pero apareciendo ahora 50 veces cada uno de los dígitos del 1 al 10 en el vector $\mathbf{x}'$. ¿Qué cambios se observan?

Tres propiedades deseables en un estimador son:

1) *Insesgadez*. Un estimador $T = T(x_1, ..., x_n)$ de un parámetro θ de la v.a. x es *insesgado* si su esperanza coincide con el valor del parámetro estimar, $E[T] = \theta$.

2) *Eficiencia*. La eficiencia de un estimador se mide por su varianza, lo ideal para un estimador insesgado es que sea lo más pequeña posible.

3) *Consistencia*. Informalmente, un estimador es *consistente* si a medida que aumenta el tamaño n de la muestra, el estimador tiende al valor del parámetro a estimar. Un estimador es consistente si es asintoticamente insesgado, $\lim_{n\to\infty} E[T] = \theta$, y con varianza asintóticamente nula, $\lim_{n\to\infty} V[T] = 0$. Por tanto, para los estimadores insesgados, basta con que su varianza tienda a cero.

Es conveniente que, si es posible, un estimador sea insesgado y entre los insesgados, que sea lo más eficiente posible. Aún fallando las dos propiedades anteriores, se le exige que las cumplan al menos asintóticamente, es decir que sea consistente.

Un estimador de α o de β se dice que es lineal, si es lineal en la muestra aleatoria $y_1, ..., y_n$, ie, del tipo $\sum \lambda_i y_i$ ($\lambda_i \in \mathbb{R}$). Se observa que $\hat{\alpha}$ y $\hat{\beta}$ son estimadores lineales, pues

$$\hat{\beta} = \sum c_i y_i, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \sum (\frac{1}{n} - c_i \bar{x}) y_i = \sum d_i y_i$$

(recordemos que x_i no son v.a. y no lo será, por tanto, $\bar{x}$).

Teorema de Gauss-Markov. *Los estimadores $\hat{\alpha}$ y $\hat{\beta}$ son los estimadores insesgados con menor posible varianza entre todos todos los estimadores lineales de α y β, respectivamente.*

No demostramos aquí este teorema, pues se hará en la situación más general de la regresión mútiple.

2.3. Inferencia

Aunque no vamos a estudiar aquí con detalle inferencia, pues también la estudiaremos con más generalidad en el tema siguiente, avanzamos algún resultado. Si el modelo estadístico que se basa en las hipótesis (H1)-(H6) está correctamente especificado (es correcto), entonces,

$$\text{la variable } y \text{ depende de } x \Leftrightarrow \beta \neq 0.$$

Por tanto, estaremos fundamentalmente interesados en contrastes con hipótesis nulas

$$\text{H0: } \beta = 0.$$

El rechazo de esta hipótesis significa que existe dependencia (variabilidad) de la variable explicada respecto a la explicativa, ie, que la explicativa realmente "explica"los cambios en la explicada. Para ello necesitamos una cantidad pivotal para β, es decir, un estadístico de la muestra $y_1, ..., y_n$ y del parámetro cuya distribución no dependa de β:

$$\frac{\hat{\beta} - \beta}{s_{\hat{\beta}}} \sim t_{n-2},$$

siendo t_{n-2} la distribución t de Student con $n-2$ g.d.l. (grados de libertad) y

$$s_{\hat{\beta}} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}, \text{ con } s^2 = \frac{1}{n-2} \sum \hat{\varepsilon}_i^2, \tag{2.15}$$

con $\hat{\varepsilon}_i := y_i - \hat{\alpha} - \hat{\beta} x_i$ (resíduos de la muestra). A

$$t_{\hat{\beta}} := \frac{\hat{\beta}}{s_{\hat{\beta}}}$$

para la muestra empírica se le llama el t-*valor* de β, $s_{\hat{\beta}}$ será el *error estándar de* $\hat{\beta}$ y s *el error estándar de la regresión*. Además, s^2 es un estimador insesgado de la varianza de la perturbación σ^2.

Recordemos que la distribución t_r es simétrica respecto al origen (ie, su densidad es una función par). Entonces si la estimación $\hat{\beta}$ del estimador de β no es pequeña, ie, $|\hat{\beta}| > c s_{\hat{\beta}}$, para un cierto valor crítico c, (¡ojo!: como es habitual en inferencia, aquí identificamos al estimador con su estimación

puntual que es un número) $\Leftrightarrow |t_{\hat{\beta}}| > c$, la hipótesis H0 será rechazada. Como en todo proceso de inferencia existe también un p-valor, cuyos valores pequeños indicarían que H0 será rechazada: es el valor crítico de la significación para H0. Naturalmente podríamos aquí también plantear intervalos de confianza para el parámetro β.

Para estudiar el intercepto α existe un test similar, también con una distribución t_{n-2}, con su correspondiente t-valor, desviación estándar y p-valor.

Ejercicio II.6. Basándonos en los datos del ejercicio II.2 de la sección 2.1 , obtener el error estándar de la regresión, la desviación estándar de la pendiente y el t-valor de la misma. Utilizando una tabla de la distribución t de Student, ¿es razonable afirmar que el precio de la viga depende realmente de su resistencia?

EJERCICIO CON ORDENADOR II.C. Hacer le ejercicio anterior mediante ordenador, obteniendo además, los p-valores de la pendiente y del intercepto.

EJERCICIO CON ORDENADOR II.D: consumo-renta. Recordemos que según Keynes el consumo es lineal respecto al ingreso. La tabla siguiente da una estimación del consumo anual medio y la renta media de los españoles durante 15 años (en miles de euros).

AÑO	RENTA	CONSUMO
1970	1959,75	1751,87
1971	2239,09	1986,35
1972	1613,84	2327,9
1973	3176,06	2600,1
1974	3921,6	3550,7
1975	4624,7	4101,7
1976	5566,02	5012,6
1977	6977,84	6360,2
1978	8542,51	7990,13
1979	9949,9	9053,5
1980	11447,5	10695,4
1981	13123,04	12093,8
1982	15069,5	12906,27
1983	16801,6	15720,1
1984	18523,5	17309,7

Estudiar si esta muestra corrobora el modelo de Keynes, obteniendo una estimación de la pendiente y del intercepto del modelo, junto con el p-valor

de la pendiente. ¿Es razonable afirmar que realmente el consumo depende de los ingresos para esta población?

Capítulo 3

Regresión lineal múltiple

En la mayor parte de los casos prácticos una sola variable explicativa no basta para que el modelo explique bien el comportamiento de la variable explicada. Por tanto, será conveniente introducir más variables explicativas. Por ejemplo, a pesar de los buenos resultados en el ejercicio del tema anterior sobre la relación ingreso-consumo de Keynes (1936), en otros experimentos se llegó a la conveniencia de introducir más variables explicativas, para explicar el consumo de las unidades familiares, como hizo Milton Friedman (1957, *hipótesis del ingreso permanente*), que consideró que también influían en el consumo las expectativas de ingreso, o Modigliani (1949, *hipótesis del ciclo de vida*), incluyendo otras variables explicativas como la riqueza de la unidad familiar.

3.1. Estimación

3.1.1. Mínimos cuadrados

Aquí partimos ya directamente de un modelo estadístico, sin pasar previamente por el modelo "matemático", pero obteniendo los coeficientes mediante mínimos cuadrados. El modelo surge de tomar muestras con varias variables explicativas, $x_2, ..., x_k$, viene definido por la ecuación

$$y_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \ i = 1, ..., n. \tag{3.1}$$

Realmente esta es una de las hipótesis del modelo estadístico (más adelante se precisarán las hipótesis). La notación es un reflejo de considerar la variable $x_1 = 1$.

La ecuación (3.1) podemos reescribirla como

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.2}$$

siendo

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{21} & \cdots & x_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{2n} & \cdots & x_{kn} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{3.3}$$

Fijémonos que el orden de los subíndices fila/columna en la matriz X no es el estándar.

A partir de la muestra empírica dada por "la nube de puntos" $(X, \mathbf{y})$ (mantenemos la notación, aunque $\mathbf{y}$ ya no es un vector aleatorio!) podemos considerar el vector de resíduos

$$\hat{\boldsymbol{\varepsilon}} = \begin{pmatrix} \hat{\varepsilon}_1 \\ \vdots \\ \hat{\varepsilon}_n \end{pmatrix} := \mathbf{y} - X\hat{\boldsymbol{\beta}}.$$

Como en el tema anterior, el objetivo es obtener una estimación $\hat{\boldsymbol{\beta}}$ del vector $\boldsymbol{\beta}$ mediante *mínimos cuadrados*, imponiendo que la suma de los cuadrados de los resíduos (la norma al cuadrado del vector $\hat{\boldsymbol{\varepsilon}}$),

$$g(\hat{\boldsymbol{\beta}}) = \sum \hat{\varepsilon}_i^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = (\mathbf{y} - X\hat{\boldsymbol{\beta}})'(\mathbf{y} - X\hat{\boldsymbol{\beta}}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'X\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'X'\mathbf{y} + \hat{\boldsymbol{\beta}}'X'X\hat{\boldsymbol{\beta}}, \tag{3.4}$$

sea mínima.

Ejercicio III.1. Demostrar que

$$\mathbf{y}'X\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}'X'\mathbf{y}.$$

Sol. Como $\mathbf{y}'X\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}'X'\mathbf{y})'$, basta demostrar que $\mathbf{y}'X\hat{\boldsymbol{\beta}}$ es una matriz simétrica, lo que es evidente al ser una matriz de orden 1 (un escalar).

Por tanto, podemos escribir (3.4) como

$$g(\hat{\boldsymbol{\beta}}) = \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'X'\mathbf{y} + \hat{\boldsymbol{\beta}}'X'X\hat{\boldsymbol{\beta}}. \tag{3.5}$$

El gradiente (jacobiano) de $g(\hat{\boldsymbol{\beta}})$ es un vector columna de dimensión n dado por

$$\left(\frac{\partial g}{\partial \hat{\beta}_i} \right) = -2X'\mathbf{y} + 2X'X\hat{\boldsymbol{\beta}}. \tag{3.6}$$

el primer sumando de la igualdad se obtiene fácilemente al derivar el sumando lineal $2\hat{\boldsymbol{\beta}}'X'\mathbf{y}$ y, aunque no lo demostramos en general, es intuitivo que el segundo sumando (que es lineal) proviene de la derivación de la parte cuadrática en (3.5).

Ejercicio III.2. Explicitar los cálculos matriciales anteriores para la regresión simple (ie, $k = 2$), verificando en particular la fórmula (3.6) para el gradiente de g y que coincide con lo obtenido en la subsección 2.1.

Igualando a cero obtenemos las llamadas *ecuaciones normales*

$$X'X\hat{\boldsymbol{\beta}} = X'\mathbf{y}, \tag{3.7}$$

es decir,

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}. \tag{3.8}$$

Estamos suponiendo que la matriz de dimensión k, $X'X$, es regular, lo cual equivale a que *el rango de la matriz X sea k* (véase ejercicio a continuación). Para esto es necesario que el número de observaciones n no sea menor que el número k de variables explicativas.

Ejercicio III.3 (de álgebra lineal).

a) Dada una matriz X, demostrar que

$$\ker(X'X) = \ker X.$$

b) Demostrar que rang $(X'X) =$ rang X.

Sol.

a) Evidentemente, $\ker(X'X) \supset \ker X$. Sea $\mathbf{v} \in \ker(X'X)$, denotando $\mathbf{w} := X\mathbf{v}$, tenemos

$$X'\mathbf{w} = \mathbf{0} \Rightarrow \mathbf{v}'X'\mathbf{w} = 0 \Rightarrow \mathbf{w}'\mathbf{w} = \mathbf{v}'X'\mathbf{w} = 0 \Rightarrow X\mathbf{v} = \mathbf{w} = \mathbf{0},$$

ie, $\mathbf{v} \in \ker X$.

b) Si X es una matriz $n \times k$, entonces

$$\dim \ \ker(X'X) + \operatorname{rang}(X'X) = k = \dim \ \ker X + \operatorname{rang}\ X,$$

y el resultado se obtiene de a).

Recordemos también que un punto crítico de una función es un punto donde se anula el gradiente de la función y que un punto crítico es no degenerado si la matriz Hessiana es regular en ese punto. Entonces un punto crítico no degenerado será un mínimo si la matriz Hessiana es definida positiva.

Para el punto crítico dado por (3.8), suponiendo que el rango de X es k se puede demostrar (aunque no lo hacemos, en general) que la matriz Hessiana

$$\left(\frac{\partial^2 g}{\partial\hat{\beta}_i \partial\hat{\beta}_j}\right) = 2X'X$$

es definida positiva ($\Leftrightarrow$ que lo sea $X'X$). Con lo cual, efectivamente, el punto obtenido es un mínimo.

Ejercicio III.4. Verificar que para la regresión simple, el cálculo del mínimo obtenido aquí coincide con lo que se obtuvo en la subsección 2.1.1, comprobando, además, que la matriz $X'X$ efectivamente es definida positiva.

Ejercicio III.5.

a) Demostrar que
$$X'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}. \tag{3.9}$$

b) Demostrar que
$$\bar{\hat{\boldsymbol{\varepsilon}}} = \frac{1}{n}\sum \hat{\varepsilon_i} = 0. \tag{3.10}$$

c) Demostrar que
$$\sum x_{ji}\hat{\varepsilon_i} = 0,\ j = 1, ..., k. \tag{3.11}$$

d) Demostrar que si $\hat{\mathbf{y}} := X\hat{\boldsymbol{\beta}}$ (variación "explicada"de la variable independiente), entonces
$$\hat{\mathbf{y}}'\hat{\boldsymbol{\varepsilon}} = 0, \tag{3.12}$$
es decir, los vectores $\hat{\mathbf{y}}$ y $\hat{\boldsymbol{\varepsilon}}$ son ortogonales.

Sol.

a)
$$X'\hat{\boldsymbol{\varepsilon}} = X'(\mathbf{y} - X(X'X)^{-1}X'\mathbf{y}) = \mathbf{0}.$$

b) Se obtiene a partir de a), considerando la primera componente del vector $X'\hat{\boldsymbol{\varepsilon}}$.

c) Se obtiene de las otras componentes de $X'\hat{\boldsymbol{\varepsilon}}$.

d)
$$(3.9) \Rightarrow \hat{\mathbf{y}}'\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\beta}}' X'\hat{\boldsymbol{\varepsilon}} = 0,$$
por a).

Una vez calculado el vector $\hat{\boldsymbol{\beta}}$, entonces la ecuación
$$y = (1\, x_2 \cdots x_k) \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots \hat{\beta}_k x_k$$
representa un plano, el *plano de regresión*, que aproxima la nube de puntos $(x_{2i}, ...x_{ki}, y_i)$, $i = 1, ..., n$ en el espacio de coordenadas $(x_2, ..., x_k, y)$.

Interpretación geométrica

Vamos a reinterpretar más geométricamente los resultados del ejercicio III.5.

Recordemos que un proyector es una matriz (que representará una aplicación lineal en un espacio euclídeo $\mathbb{R}^n$) P que es idempotente $P^2 = P$. Si P es un proyector, $I - P$ también lo es y tenemos una descomposición del espacio en suma directa del núcleo y la imagen de P (o de $I - P$), que se puede escribir como

$$\mathbb{R}^n = P(\mathbb{R}^n) \oplus (I-P)(\mathbb{R}^n) = \mathrm{Im}P \oplus \ker P = \ker(I-P) \oplus \mathrm{Im}(I-P), \quad (3.13)$$

ya que el núcleo de P (de $I - P$) es igual a la imagen de $I - P$ (de P, respectivamente) (suponemos que en álgebra se ha estudiado la suma directa de subespacios). Si P es una matriz simétrica, la proyección es ortogonal y la descomposición anterior es en suma de subespacios ortogonales; en efecto,

$\mathbf{v} \in \mathbb{R}^n \Rightarrow (P\mathbf{v})'(I-P)\mathbf{v} = \mathbf{v}'P(I-P)\mathbf{v} = 0$. Si no se dice lo contrario, todas las proyecciones que consideraremos serán ortogonales.

Observamos que si P es un proyector ortogonal, $I-P$ también lo será y ambos dan mediante (3.13) la descomposición de cualquier vector no nulo sobre los subespacios correspondientes ortogonales. Obviamente el rango de P será la dimensión del subespacio ortogonal sobre el que se proyecta $\mathrm{Im}(P) = P(\mathbb{R}^n)$, sobre el cual P es la identidad (de hecho es el autoespacio de autovalor $\lambda = 1$, el otro autoespacio de autovalor 0 es el otro sumando directo $(I-P)(\mathbb{R}^n)$).

La descomposición obtenida del método de minimos cuadrados

$$\mathbf{y} = X\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}} = P\mathbf{y} + \hat{\boldsymbol{\varepsilon}}, \tag{3.14}$$

con

$$P = X(X'X)^{-1}X',$$

es de hecho una descomposición ortogonal en el espacio $\mathbb{R}^n$. En efecto, P es un proyector, pues es simétrica y $P^2 = P$ (ejercicio fácil, teniendo en cuenta que la transpuesta de la inversa es la inversa de la transpuesta). Entonces, basta ver que

$$\hat{\boldsymbol{\varepsilon}} = M\mathbf{y}, \text{ con } M = I - P = I - X(X'X)^{-1}X', \tag{3.15}$$

que es simplemente otra forma de escribir (3.14) $\mathbf{y}$=P$\mathbf{y}$+M$\mathbf{y}$.

(3.16)

Luego lo que hace el método de mínimos cuadrados es *minimizar el módulo del vector* $\hat{\boldsymbol{\varepsilon}}$*, proyectando mediante* P *sobre el plano de dimensión* k *generado por las columnas de* $X = P(\mathbb{R}^n) := \pi_X$, (ejercicio: verificar esta afirmación, teniendo en cuenta que el rango del proyector $P = X(X'X)^{-1}X'$ es k) ese plano será el núcleo de la otra proyección M y, por tanto,

$$MX = 0. \tag{3.17}$$

Podemos escribir la descomposición (3.14) como

$$\mathbf{y} = \hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}}, \tag{3.18}$$

siendo $\hat{\mathbf{y}} = P\mathbf{y}$ *la componente explicada del vector* $\mathbf{y}$ (es decir, representan los puntos que pasan por el plano de regresión).

Está claro que rang$M = n - k$, ya que como ker M =ImP,
n =dimker M+rang$M$ =rang$P$+rang$M$.

Fijémonos que si $n = k$ (X será una matriz cuadrada) y el rango de X es k, entonces $\hat{\boldsymbol{\varepsilon}} = 0$ (es decir, el plano pasa por todos los puntos de la nube de puntos de los datos: los residuos son nulos), ya que P es la identidad (¿por qué?). Naturalmente esto era de esperar.

Un ejercicio interesante es volver sobre la regresión simple a la luz de esta interpretación geométrica.

Ejercicio III.6 (datos tomados de [2]) Dados los siguientes datos

y	x_2	x_3
3	3	5
1	1	4
8	5	6
3	2	4
5	4	6

a) Obtener el plano de regresión y el vector de resíduos.

b) Verificar que la descomposición (3.18) es ortogonal.

c) Calcular el proyector P que proyecta sobre el plano X.

d) Obtener el plano π_X.

Ayuda: para facilitar los cálculos

$$\begin{pmatrix} 5 & 15 & 25 \\ 15 & 55 & 81 \\ 25 & 81 & 129 \end{pmatrix}^{-1} = \begin{pmatrix} 26.7 & 4.5 & -8 \\ 4.5 & 1 & -1.5 \\ -8 & -1.5 & 2.5 \end{pmatrix}. \tag{3.19}$$

Solución.

a) Ante todo: el método funciona por que el rango de X es 3. Tenemos

$$X = \begin{pmatrix} 1 & 3 & 5 \\ 1 & 1 & 4 \\ 1 & 5 & 6 \\ 1 & 2 & 4 \\ 1 & 4 & 6 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 3 \\ 1 \\ 8 \\ 3 \\ 5 \end{pmatrix}, \quad X'X = \begin{pmatrix} 5 & 15 & 25 \\ 15 & 55 & 81 \\ 25 & 81 & 129 \end{pmatrix}, \tag{3.20}$$

obtenemos

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y} = \begin{pmatrix} 4 \\ 2.5 \\ -1.5 \end{pmatrix}.$$

El plano de la regresión es

$$y = 4 + 2.5x_2 - 1.5x_3.$$

b) Además,

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - X\hat{\boldsymbol{\beta}} = \begin{pmatrix} -1 \\ 0.5 \\ 0.5 \\ 0 \\ 0 \end{pmatrix},$$

de donde deducimos la descomposición

$$\begin{pmatrix} 3 \\ 1 \\ 8 \\ 3 \\ 5 \end{pmatrix} = \mathbf{y} = X\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}} = \begin{pmatrix} 4 \\ 0.5 \\ 7.5 \\ 3 \\ 5 \end{pmatrix} + \begin{pmatrix} -1 \\ 0.5 \\ 0.5 \\ 0 \\ 0 \end{pmatrix}.$$

Esta descomposición es ortogonal, ie,

$$(-1 \;\; 0.5 \;\; 0.5 \;\; 0 \;\; 0) \begin{pmatrix} 4 \\ 0.5 \\ 7.5 \\ 3 \\ 5 \end{pmatrix} = 0.$$

c) El proyector P se obtiene despues de algunos cálculos mecánicos que se repiten bastante, solo se han de calcular los elementos de la diagonal y los de encima de ella al ser simétrica:

$$P = X(X'X)^{-1}X' = \begin{pmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.7 & -0.3 & 0.2 & 0.2 \\ 0.2 & -0.3 & 0.7 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.7 & -0.3 \\ 0.2 & 0.2 & 0.2 & -0.3 & 0.7 \end{pmatrix}.$$

d) Una base de π_X está dada por

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 1 \\ 5 \\ 2 \\ 4 \end{pmatrix}, \begin{pmatrix} 5 \\ 4 \\ 6 \\ 4 \\ 6 \end{pmatrix}.$$

Coeficiente de determinación

Sea $\mathbf{1}' = (1...1)$, el vector de dimensión n con coordenadas 1. Conviene plantear la variación de las variables respecto a la media muestral $y_i - \bar{y}$ ($x_i - \bar{x}$, etc.) matricialmente a través de la matriz cuadrada de dimensión n

$$A := I - \frac{1}{n}\mathbf{1}\mathbf{1}', \text{ siendo } \mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

A actuando sobre un vector, por ejemplo $\mathbf{y}$, produce el vector de desviaciones respecto a la media,

$$A\mathbf{y} = \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}.$$

Observamos, además, que A es una matriz simétrica; de hecho es un proyector (ortogonal): $A^2 = A$ (!ejercicio!).

Aplicando A a la ecuación $\mathbf{y} = X\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}$ (descomposición ortogonal), obtenemos la descomposición de las desviaciones de y respecto a la media como una parte explicada más una parte residual,

$$A\mathbf{y} = AX\hat{\boldsymbol{\beta}} + A\hat{\boldsymbol{\varepsilon}} = AX\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}. \tag{3.21}$$

La última igualdad sale de que $A\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}}$, ya que la media muestral de $\hat{\boldsymbol{\varepsilon}}$ es nula (ecuación (3.10)).

La descomposición (3.21) también es ortogonal, es decir,

$$\hat{\boldsymbol{\beta}}' X' A' \hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\beta}}' X' A \hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\beta}}' X' \hat{\boldsymbol{\varepsilon}} = 0, \tag{3.22}$$

por (3.9). Ahora ya podemos calcular el cuadrado del módulo del vector (3.21) que será la suma de los módulos al cuadrado de cada uno de los sumandos (teorema de Pitágoras)

$$\sum (y_i - \bar{y})^2 = \mathbf{y}' A \mathbf{y} = \hat{\boldsymbol{\beta}}' X' A X \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}} : \; SST = SSE + SSR,$$

descomposición de la variación de la variable dependiente en parte explicada (por x) y parte residual.

Como en regresión simple, definimos el *coeficiente de determinación* de la regresión

$$R^2 := \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = 1 - \frac{\hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}}{\mathbf{y}' A \mathbf{y}}. \tag{3.23}$$

Observamos que la expresión que hemos obtenido en función de los resíduos y de la variación de y es la misma que la obtenida, fórmula (2.8), para la regresión simple. Extrayendo la raiz cuadrada obtendríamos el coeficiente de correlación R, que lo podemos interpretar como el coseno del ángulo entre los vectores $A\mathbf{y}$ y $AX\bar{\boldsymbol{\beta}}$.

Como en la regresión simple, el coeficiente de determinación está claramente entre 0 y 1 y es una medida de cuan bueno es el ajuste de la muestra al plano de regresión, el ajuste es bueno si es cercano a uno.

Muchos programas suelen dar también el llamado *coeficiente de determinación ajustado*

$$\bar{R}^2 := 1 - \frac{n-1}{n-k}(1 - R^2), \tag{3.24}$$

que tiene en cuenta el número de variables explicativas.

Ya comentamos que cuando $n = k$ =rang(X) (X será una matriz cuadrada) el vector de resíduos es nulo y entonces $R^2 = 1$, como era de esperar y el ajuste es perfecto: k vectores linealmente independientes generan un

plano de dimensión k, pero $\bar{R}^2$ es indeterminado. Observamos que $\bar{R}^2 < R^2$ (¡ejercicio!). La razón para ajustar el coeficiente es que al aumentar el número de variables explicativas, aumenta artificialmente el coeficiente R^2 y de esta forma se corrige ese aumento artificial. Por tanto, es conveniente tener en cuenta ese coeficiente en los casos en que la diferencia $n - k$ es pequeña: si esa diferencia es grande, como suele suceder en la práctica en donde n es mucho mayor que k, entonces ambos coeficientes dan valores cercanos. Por tanto, tomaremos como bondad del ajuste el dado por el coeficiente ajustado, aunque no hay un acuerdo completo entre los econometristas respecto a este tema...

Ejercicio III.7. Calcular el coeficiente de determinación y el coeficiente de determinación ajustado con los datos del Ejercicio III.6. ¿Es bueno el ajuste de la nube de puntos al plano de regresión?

Sol. Se obtiene directamente a partir de las fórmulas (3.23) y (3.24), teniendo en cuenta que $\hat{\boldsymbol{\varepsilon}}$ se ha obtenido en un ejercicio anterior y que SST se obtiene de los datos, $SST = 28$,

$$R^2 = 1 - \frac{1.5}{28} = 0.9464, \quad \bar{R}^2 = 1 - 2(1 - R^2) = 0.8929. \tag{3.25}$$

El plano ajusta razonablemente los datos, pues el coeficiente ajustado no está lejos de 1.

EJERCICIO CON ORDENADOR III.A. Hacer el ejercicio anterior utilizando un programa de ordenador, obteniendo también el plano de regresión. Representar gráficamente los resultados, incluyendo el plano de regresión.

3.1.2. Especificación del Modelo Estadístico

Hipótesis Las hipótesis del modelo estadístico de regresión simple se extienden de manera natural a la regresión con más variables explicativas:

(H1) *Linealidad.* Los datos se generan mediante una muestra

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

como en la ecuación (3.3), con $\boldsymbol{\beta}$ vector de constantes desconocidas y X matriz de constantes conocidas fijas, con lo cual tendrán los mismos valores en otra muestra (ie, no son va).

(H2) *No colinealidad.* La matriz X tiene rango k.

(H3) *Perturbación con media cero*:

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

es un vector aleatorio con $E[\varepsilon_i] = 0$.

(H4) *Homocedasticidad.* La varianza de la perturbación es constante en la muestra $V[\varepsilon_i] = \sigma^2$, $i = 1, ..., n$.

(H5) *No correlación.* Los pares de perturbaciones están incorrelados dos a dos $E[\varepsilon_i \varepsilon_j] = 0$, $i \neq j$, $i, j = 1, ..., n$.

Fijémonos que, como pasaba en la regresión simple, (H4) y (H5) equivalen a decir que la matriz de covarianzas $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2 I \Rightarrow \text{var}(\mathbf{y}) = \sigma^2 I$, siendo I la matriz identidad de dimensión n. La hipótesis (H2) implica que $n \geq k$ (en la práctica n suele ser bastante mayor que k) y que los vectores columna de X generan realmente un plano, que vimos que era algo necesario para obtener el vector $\boldsymbol{\beta}$ por mínimos cuadrados.

A este modelo se le llama *modelo de Gauss-Markov* de la regresión . También se suele suponer aquí que la perturbación es una distribución normal con esperanza cero

(H6)

$$\varepsilon_i \sim \text{ NID}(0, \sigma).$$

Observamos que el modelo estadístico depende de $k + 1$ parámetros: $\boldsymbol{\beta}$, σ. Ahora ya $\mathbf{y}$ es un vector aleatorio e insistimos de nuevo en que por $\hat{\boldsymbol{\beta}}$ *entenderemos tanto un estimador (va) de una muestra aleatoria, como un valor estimado de una muestra empírica concreta (un número real)*: si no se entiende esto, no se comprenderá lo que sigue (!). *El valor de ese estimador* (¡o su estimación!) *viene dado por la fórmula* (3.8) *de mínimos cuadrados.*

El estimador $\hat{\boldsymbol{\beta}}$ es insesgado

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{\varepsilon} \Rightarrow \tag{3.26}$$

$$E[\hat{\boldsymbol{\beta}}] = E[\boldsymbol{\beta}] + (X'X)^{-1}X'E[\boldsymbol{\varepsilon}] = \boldsymbol{\beta} + \mathbf{0} = \boldsymbol{\beta}, \tag{3.27}$$

por las hipótesis del modelo. Fijémonos en la potencia del álgebra de matrices para obtener de una forma tan simple este resultado.

Calculemos la matriz de covarianzas de $\hat{\boldsymbol{\beta}}$. Fijémonos que las matrices de covarianzas de un vector aleatorio $\mathbf{z}$ se puede obtener como $\mathrm{var}(\mathbf{z}) = E[(\mathbf{z} - E[\mathbf{z}])(\mathbf{z} - E[\mathbf{z}])']$, entendiendo que la esperanza de una matriz (aleatoria) es la matriz de esperanzas de cada uno de los elementos de matriz (que generaliza la esperanza de un vector aleatorio). Según esto, por (3.26) la matriz de covarianzas del estimador $\hat{\boldsymbol{\beta}}$ será

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})')] = (X'X)^{-1}X'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']X(X'X)^{-1} = \sigma^2(X'X)^{-1}, \tag{3.28}$$

por las hipótesis (H3) y (H4), recordando además que $X'X$ es simétrica, que la transpuesta de la inversa de una matriz es la inversa de la transpuesta (demostrar esto es un pequeño ejercicio de álgebra de matrices) y que $\sigma^2 I$ conmuta con cualquier matriz.

Ejercicio III.8. Explicitar los cálculos matriciales de las fórmulas (3.26) y (3.28) para $k = 2$. Comparar con lo obtenido en el tema 1.

En el contexto de la regresión, un estimador $\boldsymbol{\gamma}$ de $\boldsymbol{\beta}$ se llama lineal si es lineal en $\mathbf{y}$, ie, $\boldsymbol{\gamma} = A\mathbf{y}$, siendo A una matriz $k \times n$ (no aleatoria).Por la fórmula de mínimos cuadrados (3.8), $\hat{\boldsymbol{\beta}}$ es un estimador lineal.

Se cumple la siguiente propiedad:

Dada una matriz B de dimensión $k \times n \Rightarrow$ la matriz BB' es semidefinida positiva (3.29)

Fijémonos que se ha utilizado una propiedad similar, al demostrar que el extremo obtenido por mínimos cuadrados era un mínimo: si la matriz $B = X'$ tiene rango $k \Rightarrow BB' = X'X$ es definida positiva, que es consecuencia de (3.29), para el caso en que $\mathrm{rang}(B) = \mathrm{rang}(BB') = k$.

Ejercicio III.9. Demostrar la propiedad anterior.
(Idea: $0 \leq |B'\mathbf{v}|^2 = \mathbf{v}'BB'\mathbf{v}$.)

Teorema III.1 (Gauss-Markov). *Para cualquier estimador $\boldsymbol{\gamma}$ lineal insesgado de $\boldsymbol{\beta}$ la matriz var$(\boldsymbol{\gamma})$ − var$(\boldsymbol{\beta})$ es semidefinida positiva.*

Dem. Sea $\boldsymbol{\gamma} = A\mathbf{y} = AX\boldsymbol{\beta} + A\boldsymbol{\varepsilon}$ un estimador insesgado de $\boldsymbol{\beta}$, entonces como $AX\boldsymbol{\beta}$ es un vector no aleatorio (vector de números reales), las varianzas

de sus elementos y las covarianzas de sus elementos con cualquier otra va serán nulas (¡el alumno debe convencerse de ello!), por tanto, mediante un argumento análogo al que se hizo para obtener la matriz de covarianzas de $\hat{\boldsymbol{\beta}}$, fórmula (3.28),

$$\mathrm{var}(\boldsymbol{\gamma}) = \mathrm{var}(A\boldsymbol{\varepsilon}) = \sigma^2 AA'. \tag{3.30}$$

Entonces por (3.28),

$$\mathrm{var}(\boldsymbol{\gamma}) - \mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(AA' - (X'X)^{-1}). \tag{3.31}$$

Luego bastaría ver que la matriz $k \times k$ simétrica $AA' - (X'X)^{-1}$ es semidefinida positiva. Esto ocurre si existe una matriz $k \times n$, B, tal que

$$AA' - (X'X)^{-1} = BB'.$$

Veamos que $B = A - (X'X)^{-1}X'$ es una tal matriz

$$BB' = (A - (X'X)^{-1}X')(A' - X(X'X)^{-1}) =$$
$$AA' - (X'X)^{-1}X'A' - AX(X'X)^{-1} + (X'X)^{-1}X'X(X'X)^{-1} =$$

$$AA' - (X'X)^{-1}X'A' - AX(X'X)^{-1} + (X'X)^{-1}. \tag{3.32}$$

Ahora bien

$$\boldsymbol{\beta} = E[\boldsymbol{\gamma}] = E[A\mathbf{y}] = AE[\mathbf{y}] = AX\boldsymbol{\beta}, \text{ para todo } \boldsymbol{\beta} \Rightarrow AX = I.$$

Por tanto, de (3.32) obtenemos,

$$BB' = AA' - (X'X)^{-1},$$

como deseábamos, hemos demostrado el teorema.

En los textos en inglés se utiliza el acrónimo BLUE (Best Linear Unbiased Estimator) para referirse a esta propiedad del estimador $\hat{\boldsymbol{\beta}}$ de mínimos cuadrados. Nos dice que es el estimador lineal más eficiente del vector $\boldsymbol{\beta}$ de parámetros del modelo; aquí la eficiencia incluye también las covarianzas.

Ejercicio III.10. Demostrar que para la regresión simple con $\beta_1 = \alpha$ y $\beta_2 = \beta$ (en la notación del tema 1), la varianza de los estimadores correspondientes

por minimos cuadrados es la menor posible entre los estimadores lineales (así fue enunciado el teorema de Gauss-Markov en el capítulo 1). (Idea: los elementos de la diagonal de BB' no son negativos).

Ejercicio III.11.

a) Obtener la matriz de covarianzas var($\hat{\boldsymbol{\beta}}$) para los datos Ejercicio III.6 de la subsección 2.1.1, suponiendo que el error es $\boldsymbol{\varepsilon}_i \sim \text{NID}(0, 3)$, $i = 1, ..., 5$.

b) ¿Es esa matriz definida positiva?

c) A la vista de la matriz de covarianzas obtenida, ¿qué información obtenemos sobre la estimación de los parámetros $\hat{\boldsymbol{\beta}}$?

Sol.

a)

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix}, \quad \text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (X'X)^{-1} =$$

$$9 \begin{pmatrix} 26.7 & 4.5 & -8 \\ 4.5 & 1 & -1.5 \\ -8 & -1.5 & 2.5 \end{pmatrix} = \begin{pmatrix} 240.3 & 40.5 & -72 \\ 40.5 & 9 & -13.5 \\ -72 & -13.5 & 22.5 \end{pmatrix},$$

recordando que la matriz $(X'X)^{-1}$ ya la teníamos del Ejercicio III.6.

b) Como $\sigma^2 (X'X)^{-1}$, sabiendo que es semidefinida, necesariamente es definida positiva, pues su determinante no puede ser cero y no es negativo. Por tanto, siempre toda matriz de covarianzas del estimador $\hat{\boldsymbol{\beta}}$, var($\hat{\boldsymbol{\beta}}$), será definida positiva.

c) Se observa que los tres estimadores están correlacionados: todos los elementos de matriz fuera de la diagonal son no nulos. Además todas las varianzas (elementos de la diagonal) son también diferentes de cero. La mayor es la correspondiente al intercepto var($\hat{\beta}_1$) = 240.3, lo cual indica bastante indeterminación (error) en la estimación de ese parámetro.

Observación: hemos podido calcular la matriz de covarianzas, debido a que las variables x_{ij} son constantes (no son va) y a que nos han dado σ, pero usualmente no conocemos este parámetro.

Ahora estimamos el parámetro restante σ. A priori, al estar los resíduos $\hat{\varepsilon}_i$ muestralmente vinculados a los errores ε_i, puede intentar considerarse la varianza muestral de los resíduos

$$\frac{1}{n}\sum \hat{\varepsilon}_i^2 \tag{3.33}$$

como candidato a estimador de σ^2, pero como ocurría en la estimación de muestras simples, este estimador es sesgado. Entonces para corregir ese sesgo, lo que haremos es calcular la esperanza de $\sum \hat{\varepsilon}_i^2$. Por (3.17) y (3.15),

$$\hat{\boldsymbol{\varepsilon}} = M\mathbf{y} = M(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = M\boldsymbol{\varepsilon} \Rightarrow \tag{3.34}$$

$$E[\hat{\boldsymbol{\varepsilon}}] = 0, \quad \mathrm{var}(\hat{\boldsymbol{\varepsilon}}) = E[M\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'M'] = ME[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']M = \sigma^2 M^2 = \sigma^2 M, \tag{3.35}$$

al ser M un proyector (ortogonal).

Entonces

$$E[\sum \hat{\varepsilon}_i^2] = E[\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}] = E[\mathrm{tr}(\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}')] = \mathrm{tr}(E[\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}']) = \sigma^2 \mathrm{tr}M. \tag{3.36}$$

Pero recordando que $M = I - X(X'X)^{-1}X'$, fórmula (3.15), entonces

$$\mathrm{tr}M = \mathrm{tr}I - \mathrm{tr}(X(X'X)^{-1}X') = n - \mathrm{tr}(X'X(X'X)^{-1}) = n - k, \tag{3.37}$$

teniendo en cuenta que la traza cumple $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, para A y B tales que el número de filas de A sea igual al número de columnas de B y viceversa (esto sale de la identidad $\sum_i a_{ij}b_{ji} = \sum_j b_{ij}a_{ji}$).

Por tanto,

$$E[\sum \hat{\varepsilon}_i^2] = (n-k)\sigma^2 \Rightarrow \tag{3.38}$$

$$\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k} := s^2 \textit{ es un estimador insesgado de } \sigma^2 \tag{3.39}$$

(esta afirmación la hicimos también sin demostrarla en el caso de $k = 2$ en el tema 1). A s se le llama *error (o desviación) estándar de la regresión (o error estándar residual).*

Fijémonos que s disminuye con el tamaño de la muestra y aumenta con los resíduos, algo razonable. Un caso límite de bondad del ajuste es para $n = k = \mathrm{rang}X$, entonces $R^2 = 1$, los resíduos son cero (por (3.34), ya que

$P = I$ y $M = I - I = 0$), como era de esperar, pero no podemos decir nada sobre σ y $\bar{R}^2$ es indeterminado (como ya se comentó).

Ejercicio III.12.

a) Demostrar que la desviación típica (o estándar) del estimador $\hat{\boldsymbol{\beta}}_j$ (es decir, la raiz cuadrada de su varianza) está dada por $\sigma\sqrt{a_{jj}}$, $j = 1, ..., k$, siendo $(a_{ij}) := (X'X)^{-1}$.

b) Demostrar que $s_j := s\sqrt{a_{jj}}$ es un estimador insesgado de la desviación típica de $\hat{\beta}_j$.

c) Obtener s_2 para $k = 2$ y verificar que coincide con la expresión obtenida en la fórmula (2.15) del tema 1 (con esto interpretamos y justificamos la notación y terminología utilizada allí).

Solución. Aunque este ejercicio está muy relacionado con el ejercicio previo, vamos a hacerlo como si no tuviese relación. Los apartados a) y b) son consecuencia inmediata de las definiciones y de la teoría anterior:

a) Por definición, la diagonal de la matriz de covarianzas del vector aleatorio $\hat{\boldsymbol{\beta}}$, $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(X'X)^{-1} = \sigma^2(a_{ij})$, nos da el vector de varianzas (¡que no es un vector aleatorio!):

$$\begin{pmatrix} V[\hat{\beta}_1] \\ \vdots \\ V[\hat{\beta}_k] \end{pmatrix} = \text{diag}(\text{var}(\hat{\boldsymbol{\beta}})) = \text{diag}(\sigma^2(a_{ij})) = \sigma^2 \begin{pmatrix} a_{11} \\ \vdots \\ a_{kk} \end{pmatrix}.$$

Como la desviación típica de $\hat{\beta}_j$ es la raiz cuadrada de su varianza, su valor es $\sigma\sqrt{a_{jj}}$.

b)

$$s^2 \text{ es estimador insesgado de } \sigma^2 \Rightarrow s \text{ es estimador insesgado de } \sigma \Rightarrow$$

$$s\sqrt{a_{jj}} \text{ es estimador insesgado de } \sigma\sqrt{a_{jj}}.$$

c) Hemos de demostrar que para la regresión simple con $k = 2$,

$$s_2 = s\sqrt{a_{22}} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}} \tag{3.40}$$

(β_2 se denotaba como β en el tema 1). Para la regresión simple

$$X'X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & (\sum x_i)^2 \end{pmatrix} \Rightarrow (X'X)^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} (\sum x_i)^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} = (a_{ij}).$$

Tenemos que calcular

$$a_{22} = \frac{n}{n\sum x_i^2 - (\sum x_i)^2} = \frac{1}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

(de hecho su raíz cuadrada). Observando la fórmula (3.40), con un cálculo similar al realizado tantas veces para la varianza muestral,

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum x_i^2 - n(\frac{1}{n}\sum x_i)^2 \Rightarrow a_{22} = \frac{1}{\sum (x_i - \bar{x})^2},$$

es decir, hemos demostrado (3.40).

Observamos que también podríamos obtener s_1.

Al estimador $s_j := s\sqrt{a_{jj}}$ (que también se escribe $s_{\hat{\beta}_j}$) de la desviación típica de $\hat{\beta}_j$ se le llama *error estándar (o típico) del coeficiente estimado* $\hat{\beta}_j$. Observamos que, por tanto, se ha obtenido en realidad un estimador insesgado, $s^2 a_{jj}$ de $V[\hat{\beta}_j] = \sigma^2 a_{jj}$: elementos diagonales de la matriz de covarianzas var($\hat{\boldsymbol{\beta}}$)). Ya se dijo anteriormente que normalmente no se podia calcular esa matriz, debido a que no conocemos el parámetro σ^2, pero ahora al tener un estimador adecuado, s^2, lo podemos estimar y obtener por tanto, a partir de los datos, una estimación $s^2(X'X)^{-1}$ de la matriz var($\hat{\boldsymbol{\beta}}$) = $\sigma^2(X'X)^{-1}$.

Ejercicio III.13 (continuación del ejercicio III.6.) Calcular s y s_j, $j = 1, 2, 3$ para los datos del ejercicio III.6. Según esto, ¿es razonable que $\sigma = 3$, como se ha dicho en un ejercicio anterior?

EJERCICIO CON ORDENADOR III.B. Hacer el ejercicio anterior con ordenador verificando que, aparte de errores de redondeo, se obtiene lo mismo.

Ejercicio III.14. Demostrar que

$$\frac{1}{\sigma^2}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \sim \chi^2_{n-k}. \tag{3.41}$$

Sol. Por (3.34), teniendo en cuenta que M es un proyector,

$$\frac{1}{\sigma^2}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \frac{\boldsymbol{\varepsilon}'}{\sigma} M \frac{\boldsymbol{\varepsilon}}{\sigma}.$$

Ahora bien,

$$\frac{\varepsilon_i}{\sigma} \sim \mathrm{NID}(0,1),$$

al estandarizar ε_i. El resultado se sigue del Ejercicio I.3 de la Introducción, recordando que M es un proyector de rango $n-k$.

Resumimos las propiedades demostradas en relación con la estimación de σ:

Teorema III.2.

$$\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k} := s^2 \textit{ es un estimador insesgado de } \sigma^2 \tag{3.42}$$

$$\textit{La matriz } s^2(X'X)^{-1} \textit{ es un estimador insesgado de } \ \mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2((X'X)^{-1} \tag{3.43}$$

$$s^2 a_{jj} \textit{ es un estimador insesgado de } V[\hat{\beta}],\ (X'X)^{-1} = (a_{ij}) \tag{3.44}$$

$$\frac{1}{\sigma^2}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \sim \chi^2_{n-k} \tag{3.45}$$

Interpretación de los coeficientes estimados (céteris páribus)

En el modelo de regresión múltiple

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

(modelo poblacional, llamado a veces *modelo verdadero*, que es desconocido y tratamos de aproximar), aproximado por el plano de regresión

$$y = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k,$$

podemos considerar el efecto que tendría sobre la variable y el hecho de que solo variara una variable, por ejemplo x_2, *manteniéndose el resto de variables explicativas fijas*. El hecho de que se mantengan el resto de variables explicativas fijas se designa como *céteris páribus* (del latín: "siendo las demás cosas igual") y se habla del efecto céteris páribus de la variable x_2 sobre y. Usualmente en econometría no podemos hacer esto experimentalmente, es decir, tomar muestras en donde todas las variables salvo una sean constantes, pero *la regresión múltiple permite simular el experimento*. Si la variable x_2 aumenta en Δx_2 unidades, por céteris páribus su influencia sobre la variable dependiente y será

$$\Delta\hat{y} = \hat{\beta}_2\Delta x_2 + \sum_{j=3}^{k}\hat{\beta}_j\Delta x_j,\ \Delta x_j = 0,\ j = 3, ..., k \Rightarrow \Delta\hat{y} = \hat{\beta}_2\Delta x_2.$$

Matemáticamente esto es obvio, lo interesante está en cómo se interpreta en las aplicaciones.

Ejercicio III.15 (sacado de [4]). A partir de 526 observaciones sobre trabajadores de una cierta base de datos (en USA en 1977) se llegó al siguiente plano de regresión

$$y = 0.284 + 0.92x_2 + 0.0041x_3 + 0.22x_4,$$

siendo x_2 =años de educación, x_3 =años de experiencia, x_4 =años de antiguedad en la empresa, $y = \log z$, z =salario (en dólares/hora). Calcular en qué porcentaje aumenta el salario/hora un incremento de un año en educación, para trabajadores con la misma antiguedad y experiencia (céteris páribus).

Solución. El porcentaje que se busca es

$$100\frac{\Delta z}{z_0} = 100\frac{e^{y_1} - e^{y_0}}{e^{y_0}} = 100(e^{\Delta y} - 1),$$

siendo z_0 en salario/hora inicial e z_1 el final, pero $\Delta x_2 = 1$ año, luego

$$e^{\Delta y} = e^{0.92\Delta x_2} = e^{0.92} \simeq 2.51 \Rightarrow 100\frac{\Delta z}{z_0} = 151$$

La conclusión es que el modelo predice que dos trabajadores con la misma antiguedad y experiencia, pero uno de ellos con un año más de educación que

el otro, el primero ganará aproximadamente un 151 % más que el segundo. Pero fijémonos que es razonable pensar que la antiguedad o la experiencia pueden afectar a la educación (ie, que exista correlación), lo cual indica que el resultado de quitar esas variables del análisis de la regresión (ie, hacer una regresión simple entre el salario y la educación) no es un efecto tipo céteris páribus, esto es justo lo que vamos a estudiar a continuación.

3.1.3. Quitar o añadir variables explicativas

Dada la nube de datos $(X, \mathbf{y})$, del modelo estadístico que tratamos de estimar

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.46}$$

siendo X una matriz $n \times k$, veamos cómo afecta a la estimación el hecho de quitar variables explicativas del modelo. Por ejemplo, supongamos que quitamos las últimas g variables $x_{k-g+1}, ..., x_k$, es decir, estamos considerando como modelo estadístico

$$\mathbf{y} = X_1\boldsymbol{\beta_1} + \boldsymbol{\varepsilon}_R, \tag{3.47}$$

descomponiendo $X = (X_1\, X_2)$ por columnas en dos submatrices X_1, X_2 de dimensiones $n \times (k-g)$ y $n \times g$, respectivamente. El vector $\boldsymbol{\beta_1}$ es el vector de dimensión $k-g$ de coordenadas $\beta_1, ..., \beta_{k-g}$

$$\boldsymbol{\beta_1} := \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{k-g} \end{pmatrix}.$$

El modelo (3.47) es el llamado *modelo restringido* (de ahí que en la perturbación pongamos un subíndice R), en oposición al modelo completo (3.46), que ahora podemos escribir como

$$\mathbf{y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} = (X_1\, X_2)\begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{pmatrix} + \boldsymbol{\varepsilon}, \tag{3.48}$$

$$\boldsymbol{\beta_1} := \begin{pmatrix} \beta_1 \\ \vdots \\ \\ \beta_{k-g} \end{pmatrix}, \quad \boldsymbol{\beta_2} := \begin{pmatrix} \beta_{k-g+1} \\ \vdots \\ \\ \beta_k \end{pmatrix}.$$

Los datos del modelo restringido $(X_1, \mathbf{y})$ son los mismos que los del modelo completo, pero quitando las columnas de la matriz X_2. Por tanto, *no hemos cambiado el experimento mediante el que se obtienen los datos*, solo hemos dejado de considerar las variable $x_{g-k+1}, ..., x_k$ como explicativas, ie, hemos modificado el modelo de partida.

Si denotamos por

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} \text{ (vector } k \text{ dimensional)}, \quad \hat{\boldsymbol{\beta}}_R \text{ (vector } k-g \text{ dimensional)},$$

a la estimación por mínimos cuadrados de $\boldsymbol{\beta}$ o $\boldsymbol{\beta_1}$ en el modelo completo o restringido, respectivamente, ¿cómo están relacionados esos estimadores?

De lo datos $(X_1\ \mathbf{y})$, obtenemos la estimación de $\boldsymbol{\beta}_1$

$$\hat{\boldsymbol{\beta}}_R = (X_1'X_1)^{-1}X_1'\mathbf{y}, \tag{3.49}$$

con vector de resíduos $\hat{\boldsymbol{\varepsilon}}_R$. Como el plano de regresión generado por las columnas de X_1 es ortogonal al vector de resíduos,

$$X_1'\hat{\boldsymbol{\varepsilon}}_R = 0. \tag{3.50}$$

Para el modelo completo tenemos la descomposición ortogonal

$$\mathbf{y} = X\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}} = (X_1\, X_2)\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} + \hat{\boldsymbol{\varepsilon}} = X_1\hat{\boldsymbol{\beta}}_1 + X_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}}, \tag{3.51}$$

con

$$X'\hat{\boldsymbol{\varepsilon}} = 0 \Leftrightarrow X_1'\hat{\boldsymbol{\varepsilon}} = 0, \quad X_2'\hat{\boldsymbol{\varepsilon}} = 0 \tag{3.52}$$

por ortogonalidad.

Entonces, teniendo en cuenta (3.49), aplicando $(X_1'X_1)^{-1}X_1'$ a (3.51),

$$\hat{\boldsymbol{\beta}}_R = (X_1'X_1)^{-1}X_1'X_1\hat{\boldsymbol{\beta}}_1 + (X_1'X_1)^{-1}X_1'X_2\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_1 + Q\hat{\boldsymbol{\beta}}_2, \tag{3.53}$$

siendo Q la matriz $(k-g)\times g$

$$Q = (X_1'X_1)^{-1}X_1'X_2. \tag{3.54}$$

Podemos interpretar la ecuación (3.54) como una regresión intermedia entre las variables regresoras del modelo completo, con datos (X_1, X_2) (aquí no

hay modelo estadístico, solo modelo matemático, al no ser X_2 una matriz de va), regresando las columnas de X_2 respecto a las columnas de X_1. Es decir, las columnas de la matriz Q nos dan los coeficientes obtenidos por mínimos cuadrados de regresión de las variables $x_{k-g-1}, .., x_k$ respecto a las otras variables explicativas $x_2, ..., x_{k-g}$.

Por tanto, la ecuación (3.53) nos indica que *el efecto de las variables explicativas $x_2, ..., x_{k-g}$ sobre la variable independiente y en el modelo restringido tiene dos componentes*: las estimaciones de la pendiente en el modelo completo ($\hat{\boldsymbol{\beta}}_1$, céteris paribus: mateniendo fijas las otras variables) y la influencia de la relación lineal Q estimada (correlación) entre las variables $x_2, ..., x_{k-g}$ y $x_{k-g-1}, ..., x_k$, junto con su propio efecto céteris páribus en la regresión completa ($\hat{\boldsymbol{\beta}}_2$). En el caso improbable que $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ (ie, que las variables $x_{k-g-1}, ..., x_k$ realmente no sean explicativas en la muestra), o si no se detecta correlación entre los conjuntos de variables explicativas $x_{k-g-1}, .., x_k$ y $x_2, .., x_{k-g}$ ($Q = 0 \Leftrightarrow X_1$ y X_2 ortogonales, $X_1'X_2 = 0$), entonces ambas estimaciones $\hat{\boldsymbol{\beta}}_1$ y $\hat{\boldsymbol{\beta}}_R$ serán iguales. Podemos interpretar, por tanto, la ecuación (3.53) como el hecho de que aunque dejemos de considerar las variables $x_{k-g-1}, ..., x_k$ como explicativas, usualmente *seguirán estando ahí, influyendo indirectamente sobre la estimación de los parámetros.*

Observamos aquí que un caso extremo posible de modelo restringido es cuando quitamos todas las variables explicativas $x_2, ..., x_k$, estimando solamente el intercepto β_1. Volveremos sobre este modelo después del ejercicio siguiente.

Ejercicio III.16. Con los datos del Ejercicio III.6:

a) Si consideramos como modelo restringido el de las variables x_2, y, obtener la estimación de β_1 y β_2 para dicho modelo y verificar que se cumple la ecuación (3.53), interpretando dicha ecuación.

b) Si consideramos como modelo restringido el de las variables x_3, y, obtener la estimación de β_1 y β_3 para dicho modelo, y verificar que se cumple la ecuación (3.53).

c) Si consideramos como modelo restringido el de la variable y solamente (ie, quitamos las dos variables explicativas), obtener la estimación del intercepto β_1 para dicho modelo y verificar que se cumple la ecuación (3.53).

Sol. Resolveremos los apartados a) y c), pues b) es similar a a).

a) Los datos son

$$X = (X_1\, X_2) = \begin{pmatrix} 1 & 3 & 5 \\ 1 & 1 & 4 \\ 1 & 5 & 6 \\ 1 & 2 & 4 \\ 1 & 4 & 6 \end{pmatrix}, \quad X_1 = \begin{pmatrix} 1 & 3 \\ 1 & 1 \\ 1 & 5 \\ 1 & 2 \\ 1 & 4 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 5 \\ 4 \\ 6 \\ 4 \\ 6 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 3 \\ 1 \\ 8 \\ 3 \\ 5 \end{pmatrix}. \tag{3.55}$$

Entonces la estimación para el modelo restringido es

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}_R = (X_1'X_1)^{-1}X_1'\mathbf{y} = \begin{pmatrix} 5 & 15 \\ 15 & 55 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 3 & 1 & 5 & 2 & 4 \end{pmatrix} \begin{pmatrix} 3 \\ 1 \\ 8 \\ 3 \\ 5 \end{pmatrix} = \begin{pmatrix} -0.8 \\ 1.6 \end{pmatrix}. \tag{3.56}$$

Verifiquemos la relación (3.53), que aquí es

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}_R = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} + Q(\hat{\beta}_3), \quad \text{con } \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 2.5 \\ -1.5 \end{pmatrix}. \tag{3.57}$$

Obtenemos la matriz Q (aquí es un vector) regresando la variable x_3 sobre x_2,

$$Q = (X_1'X_1)^{-1}X_1'X_2 = \begin{pmatrix} 5 & 15 \\ 15 & 55 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 3 & 1 & 5 & 2 & 4 \end{pmatrix} \begin{pmatrix} 5 \\ 4 \\ 6 \\ 4 \\ 6 \end{pmatrix} = \begin{pmatrix} 3.2 \\ 0.6 \end{pmatrix} \tag{3.58}$$

$$\Rightarrow Q(\hat{\beta}_3) = \begin{pmatrix} 3.2 \\ 0.6 \end{pmatrix}(-1.5) = \begin{pmatrix} -4.8 \\ -0.9 \end{pmatrix}. \tag{3.59}$$

Se verifica la relación pedida:

$$\begin{pmatrix} -0.8 \\ 1.6 \end{pmatrix} = \begin{pmatrix} 4 \\ 2.5 \end{pmatrix} + \begin{pmatrix} -4.8 \\ -0.9 \end{pmatrix}.$$

c) Aquí

$$X_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 3 & 5 \\ 1 & 4 \\ 5 & 6 \\ 2 & 4 \\ 4 & 6 \end{pmatrix}.$$

Los cálculos son muy simples en el modelo restringido, que solo tiene intercepto,

$$\hat{\beta}_{1R} = (X_1'X_1)^{-1}X_1'\mathbf{y} = \bar{y},$$

y "el plano de regresión"será simplemente

$$y = \bar{y}, \text{ es decir } y = 4.$$

La matriz Q se obtiene también muy fácilmente, pues es un vector fila dado por las medias muestrales de las variables explicativas (¡el alumno debe verificar que es así!), en nuestro caso

$Q = (\bar{x}_2\ \bar{x}_3) = (3\ 5)$.

La relación (3.53) será

$$\bar{y} = \hat{\beta}_1 + (\bar{x}_2\, \bar{x}_3) \begin{pmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \hat{\beta}_1 + \hat{\beta}_2\bar{x}_2 + \hat{\beta}_3\bar{x}_3,$$

que nos dice simplemente que, como en el caso de la regresión simple, en el modelo completo el plano de regresión pasa por el punto de las medias muestrales. En nuestro caso,

$$4 = 4 + 2.5 \cdot 3 - 1.5 \cdot 5 = 4 + 0,$$

(el hecho de que el intercepto $\hat{\beta}_1 = 4$ coincida con la media muestral $\bar{y}$ es algo casual de este ejemplo: ¡no pasa en general!).

En el apartado c) del ejercicio se ha considerado el modelo restringido omitiendo todas las variables explicativas, que en el modelo estadístico sería

$$y = \beta_1 + \varepsilon. \tag{3.60}$$

Este modelo se llama *modelo constante*, y hemos visto que la regresión da la recta constante

$$y = \hat{\beta}_1 = \bar{y}, \tag{3.61}$$

es decir, aproximamos el conjunto de puntos $\mathbf{y}$ por su media muestral. Como los resíduos aquí son $\hat{\varepsilon}_i = y_i - \bar{y}$, por (3.23),

$$R^2 = 1 - \frac{\sum \hat{\varepsilon}_i^2}{\sum (y_i - \bar{y})^2} = 0. \tag{3.62}$$

A pesar de su simpleza, este modelo jugará un papel más adelante en inferencia (*test de significación global*).

EJERCICIO CON ORDENADOR III.C. Hacer el ejercicio anterior con R Commander.
Idea: en el programa podemos decidir cúales son las variables explicativas y la explicada. Además, para c) se puede introducir una variable explicativa ficticia x con datos nulos, definida como $x = 0 * y$: "modificar variables conjunto de datos activo-¿calcular una nueva variable...", y se hace la regresión de x sobre y y el programa calcula la recta de regresión constante, a pesar de que no se cumple la condición del rango.

Veamos ahora la relación entre los resíduos del modelo completo $\hat{\boldsymbol{\varepsilon}}$ y del modelo restringido $\hat{\boldsymbol{\varepsilon}}_R$. Si $P_1 = X_1(X_1'X_1)^{-1}X_1'$ es el proyector $\mathbb{R}^n \to \mathbb{R}^n$ sobre el espacio generado por las columnas de X_1, $M_1 = I - P_1$ el proyector sobre los resíduos $\hat{\boldsymbol{\varepsilon}}_R$,

$$\hat{\boldsymbol{\varepsilon}}_R = M_1\mathbf{y} = M_1(X_1\hat{\boldsymbol{\beta}}_1 + X_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}}) = M_1X_2\hat{\boldsymbol{\beta}}_2 + M_1\hat{\boldsymbol{\varepsilon}}, \tag{3.63}$$

ya que $M_1X_1 = 0$, al proyectar M_1 ortogonalmente al subespacio generado por las columnas de X_1. Ahora bien, por (3.52),

$$X_1'\hat{\boldsymbol{\varepsilon}} = \mathbf{0} \Rightarrow M_1\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}} \Rightarrow \hat{\boldsymbol{\varepsilon}}_R = M_1X_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}}. \tag{3.64}$$

Vemos que si

$$\hat{\boldsymbol{\beta}}_2 = \mathbf{0} \Rightarrow \hat{\boldsymbol{\varepsilon}}_R = \hat{\boldsymbol{\varepsilon}},$$

es decir, si muestralmente no se detectan las variables que hemos quitado como explicativas, entonces (¡como era de esperar!) ambos resíduos serán iguales.

La fórmula (3.64) nos permite comparar la suma de los errores cuadráticos de los resíduos, $\hat{\boldsymbol{\varepsilon}}'_R\hat{\boldsymbol{\varepsilon}}_R$ y $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$,

$$\hat{\boldsymbol{\varepsilon}}'_R\hat{\boldsymbol{\varepsilon}}_R = (\hat{\boldsymbol{\beta}}'_2X'_2M'_1 + \hat{\boldsymbol{\varepsilon}}')(M_1X_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}}). \tag{3.65}$$

Ahora bien,

$$X'_2M'_1\hat{\boldsymbol{\varepsilon}} = X'_2\hat{\boldsymbol{\varepsilon}} - X'_2X_1(X'_1X_1)^{-1}X'_1\hat{\boldsymbol{\varepsilon}} = \mathbf{0}(\Rightarrow \hat{\boldsymbol{\varepsilon}}'M_1X_2 = \mathbf{0}),$$

ya que por (3.52) $X'_1\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$ y $X'_2\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$. Por tanto,

$$\hat{\boldsymbol{\varepsilon}}'_R\hat{\boldsymbol{\varepsilon}}_R = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} + (\hat{\boldsymbol{\beta}}'_2X'_2)M_1(X_2\hat{\boldsymbol{\beta}}_2). \tag{3.66}$$

Finalmente, como M_1 es una matriz semidefinida positiva (¿por qué?),

$$\hat{\boldsymbol{\varepsilon}}'_R\hat{\boldsymbol{\varepsilon}}_R - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = (\hat{\boldsymbol{\beta}}'_2X'_2)M_1(X_2\hat{\boldsymbol{\beta}}_2) \geq 0,$$

es decir, *los errores cuadráticos del modelo restringido no son menores que los del modelo completo*, algo razonable: es de esperar que el modelo completo aproxime mejor los datos que el restringido.

Ejercicio III.17. Con los datos del Ejercicio III.6, considerando como modelo restringido el de las variables x_2, y:

a) Obtener $\hat{\boldsymbol{\varepsilon}}_R$ y $\hat{\boldsymbol{\varepsilon}}$, verificando que se cumple la ecuación (3.64).

b) Calcular $\hat{\boldsymbol{\varepsilon}}'_R\hat{\boldsymbol{\varepsilon}}_R - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$.

Sol.

$$\hat{\boldsymbol{\varepsilon}} = \begin{pmatrix} -1 \\ 0.5 \\ 0.5 \\ 0 \\ 0 \end{pmatrix} \text{ (ejercicio III.6)} , \quad \hat{\boldsymbol{\varepsilon}}_R = \begin{pmatrix} -1 \\ 0.2 \\ 0.8 \\ 0.6 \\ -0.6 \end{pmatrix}, \quad \hat{\boldsymbol{\varepsilon}}'_R\hat{\boldsymbol{\varepsilon}}_R - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = 0.9.$$

EJERCICIO CON ORDENADOR III.D. Hacer el apartado a) del ejercicio anterior con ordenador.

Si *suponemos que el verdadero modelo es el completo* (3.48), ¿qué consecuencias tiene sobre la esperanza y las varianzas de los estimadores que consideremos el modelo restringido (3.47)?

$$\mathbf{y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \Rightarrow \tag{3.67}$$

$$\hat{\boldsymbol{\beta}}_R = (X_1'X_1)^{-1}X_1'\mathbf{y} = \boldsymbol{\beta}_1 + (X_1'X_1)^{-1}X_1'X_2\boldsymbol{\beta}_2 + (X_1'X_1)^{-1}X_1'\boldsymbol{\varepsilon} \Rightarrow \tag{3.68}$$

$$E[\hat{\boldsymbol{\beta}}_R] = \boldsymbol{\beta}_1 + (X_1'X_1)^{-1}X_1'X_2\boldsymbol{\beta}_2 = \boldsymbol{\beta}_1 + Q\boldsymbol{\beta}_2, \tag{3.69}$$

coherente con el resultado muestral (3.53). Por tanto, el estimador $\hat{\boldsymbol{\beta}}_R$ del modelo restringido no es insesgado, sino que tiene un sesgo $Q\beta_2$, llamado *sesgo de las variables omitidas*. Naturalmente, si el verdadero modelo es el restringido, $\boldsymbol{\beta}_2 = \mathbf{0}$, el sesgo desaparece.

A partir de (3.68) y (3.69) se obtiene

$$\hat{\boldsymbol{\beta}}_R - E[\hat{\boldsymbol{\beta}}_R] = (X_1'X_1)^{-1}X_1'\boldsymbol{\varepsilon}, \tag{3.70}$$

con lo que obtenemos la matriz de covarianzas

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_R) = E[(\hat{\boldsymbol{\beta}}_R - E[\hat{\boldsymbol{\beta}}_R])(\hat{\boldsymbol{\beta}}_R - E[\hat{\boldsymbol{\beta}}_R])'] = \sigma^2(X_1'X_1)^{-1}. \tag{3.71}$$

Podríamos demostrar, pero no lo haremos, que la diferencia $\mathrm{var}(\hat{\boldsymbol{\beta}}_1) - \mathrm{var}(\hat{\boldsymbol{\beta}}_R)$ es semidefinida positiva, lo que significa que al considerar el modelo restringido su varianza es, en general menor, es decir *la eficiencia del modelo restringido es mayor que la del modelo completo*. Por tanto,*si el efecto de omitir las variables $x_{k-g-1}, ..., x_k$ es poco significativo, aunque aumente el sesgo, al aumentar también su precisión (menor error en la estimación), sería razonable descartarlas del modelo*. Para poder tomar este tipo de decisiones necesitaremos inferencia.

También se cumple que si el verdadero modelo es el restringido (3.47), pero trabajamos con el modelo completo (3.48) (es decir, las variables $x_{k-g+1}, ..., x_k$ son redundantes), entonces perdemos eficiencia en la estimación de los parámetros β_i. Por tanto, interesa eliminar las variables redundantes.

Ejercicio III.18. Con los datos del Ejercicio III.6, considerando como modelo restringido el de las variables x_2, y, verificar que $\mathrm{var}(\hat{\boldsymbol{\beta}}_1) - \mathrm{var}(\hat{\boldsymbol{\beta}}_R)$ es

semidefinida positiva. (Observación: aunque no conozcamos el valor de σ^2 es posible obtener la verificación pedida).

Resumiendo los resultados de esta subsección:

Teorema III.3. *Sean $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$ las estimaciones de los parámetros del modelo completo,* (3.48)*, y $\hat{\boldsymbol{\beta}}_R$ las del modelo restringido,* (3.47)*, siendo $\hat{\boldsymbol{\varepsilon}}$, $\hat{\boldsymbol{\varepsilon}}_R$ sus resíduos respectivos. Entonces:*

1)
$$\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}}_1 + Q\hat{\boldsymbol{\beta}}_2, \quad Q := (X_1'X_1)^{-1}X_1'X_2. \tag{3.72}$$

2)
$$\hat{\boldsymbol{\varepsilon}}_R = M_1X_2\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\varepsilon}} \tag{3.73}$$

3)
$$\hat{\boldsymbol{\varepsilon}}_R'\hat{\boldsymbol{\varepsilon}}_R - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\beta}}_2'X_2'M_1X_2\hat{\boldsymbol{\beta}}_2 \geq 0 \tag{3.74}$$

4)
$$E[\hat{\boldsymbol{\beta}}_R] = \boldsymbol{\beta}_1 + Q\boldsymbol{\beta}_2, \quad \text{var}(\hat{\boldsymbol{\beta}}_R) = \sigma^2(X_1'X_1)^{-1} \tag{3.75}$$

5)
$$\text{var}(\hat{\beta}_1) - \text{var}(\hat{\boldsymbol{\beta}}_R) \quad \textit{es semidefinida positiva.} \tag{3.76}$$

3.2. Inferencia

3.2.1. Test t

Este test estudia la significación de los coeficientes individuales β_j, $j = 2, ..., k$ de la regresión, es decir, su relevancia. Lo que se desea estudiar es el contraste

$$\text{H0}: \ \beta_j = 0, \quad \text{H1}: \ \beta_j \neq 0. \tag{3.77}$$

Por tanto, rechazar H0 implica que la variable y depende significativamente de la variable x_j, es decir, no deberíamos descartarla del modelo. Vamos a obtener un estadístico (cantidad pivotal) para este contraste, para ello necesitamos saber primero qué tipo de distribución sigue el estimador $\hat{\boldsymbol{\beta}}$.

Recordemos (3.26), (3.27) y (3.28):

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{\varepsilon}, \quad E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}, \quad \text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(X'X)^{-1} \Rightarrow \tag{3.78}$$

el vector aleatorio $\hat{\boldsymbol{\beta}}$ *es normal* (por (1.4)) *con* $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(XX')^{-1})$. (3.79)

Entonces si $(a_{ij}) = (XX')^{-1}$,

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 a_{jj}) \Rightarrow \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{a_{jj}}} \sim N(0,1), \tag{3.80}$$

al estandarizar.

Ahora, como siempre, el problema es que no conocemos σ, luego hemos de sustituirla por su estimación insesgada s y esto nos llevará a una distribución t, en efecto,

$$\frac{\hat{\beta}_j - \beta_j}{s_j} = \frac{\hat{\beta}_j - \beta_j}{s\sqrt{a_{jj}}} = \frac{\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{a_{jj}}}}{\frac{s}{\sigma}}, \tag{3.81}$$

pero por la definición de s^2 (3.39), y por (3.41)

$$s^2 := \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k}, \quad \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\sigma^2} \sim \chi^2_{n-k} \Rightarrow \frac{s^2}{\sigma^2} = \frac{\chi^2_{n-k}}{n-k} \Rightarrow \frac{\hat{\beta}_j - \beta_j}{s_j} \sim t_{n-k} := t(n-k), \tag{3.82}$$

por la definición de la distribución $t(n-k)$. Aquí escribiremos $t(n-k)$ cuando pueda existir confusión con los t-valores (ver más adelante).

Para el desarrollo anterior sea válido, tiene que ocurrir que los estadísticos $\hat{\beta}_j$ y $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ sean independientes, admitiremos esto sin demostración. Luego

Teorema III.4. *Una cantidad pivotal para el parámetro* β_j *es*

$$\frac{\hat{\beta}_j - \beta_j}{s_j} \sim t(n-k). \tag{3.83}$$

El t-valor del parámetro $\hat{\beta}_j$ (o de la variable x_j) es la estimación para $\beta_j = 0$ de la cantidad pivotal anterior,

$$t_j := \frac{\hat{\beta}_j}{s_j} \sim t_{n-k}, \quad j = 2, ..., k. \tag{3.84}$$

Como en el caso de la regresión simple,*si t_j es significativamente distinto de cero, entonces rechazamos H0 al caer en la región crítica, admitimos que x_j tiene un efecto significativo sobre la variable explicada y, por tanto, no podemos descartarla del modelo. En vez del t-valor podemos considerar el p-valor: valores pequeños indican que debemos rechazar H0.* Usualmente, se fija una significación del 5 % en este test, es decir, p-valores por debajo de este hacen que se rechace H0. Si n es grande ($n - k > 30$, $n - k$ son los llamados grados de libertad del test t), podemos aproximar t_{n-k} por la normal estandarizada y un t-valor mayor que 2 hace que rechacemos H0 al 5 %, siendo el test de dos colas (¡verificarlo con una tabla o con un programa informático!). Se observa que el t valor depende tanto del valor estimado $\hat{\beta}_j$ como del error estándar $s_j = s\sqrt{a_{jj}}$, de modo que si ese error (que, al ser una estimación de la desviación típica de $\hat{\beta}_j$, nos indica indeterminación en la estimación de ese parámetro) es grande entonces es más probable que tengamos que aceptar que la variable x_j no es significativa, $j = 2, ..., k$.

Ejercicio III.19. Con los datos del ejercicio III.6, obtener los t-valores. ¿Es significativa la variable x_2?, ¿y la variable x_3?

Sol. $t_2 = 2.887$, $t_3 = -1.095$. Al 5 %, con la tabla del estadístico $t(2)$, en ambos casos no podemos rechazar H0 y cada una de esas variables (céteris páribus) no es significativa (recordemos que es un test de dos colas). De hecho el p-valor de x_2 está algo por encima de 0.1 y el de x_3 claramente por encima de 0.2.

EJERCICIO CON ORDENADOR III.E. Hacer el ejercicio anterior con ordenador.

3.2.2. Test F

El F-test es una generalización del t-test. Trata de resolver el problema de la significación de los modelos restringidos, estudiados en la subsección 3.1.3, respecto al modelo completo. Recordemos que el modelo restringido es

$$\mathbf{y} = X_1\boldsymbol{\beta_1} + \boldsymbol{\varepsilon}_R, \tag{3.85}$$

obtenido al quitar las g variables, $x_{k-g+1}, ..., x_k$, del modelo completo

$$\mathbf{y} = X_1\boldsymbol{\beta}_1 + X_1\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}. \tag{3.86}$$

El problema de significación es:

$$\text{H0}: \boldsymbol{\beta}_2 = \mathbf{0}, \quad \text{H1}: \boldsymbol{\beta}_2 \neq \mathbf{0}. \tag{3.87}$$

Fijémonos que H0 es explícitamente $\beta_{k-g+1} = \cdots = \beta_k = 0$. Si se rechaza H0 entonces el test sugiere la conveniencia de mantener el modelo completo *frente al restringido*: el test solo compara ambos modelos. Fijémonos que H1 significa que, al menos, alguno de los g parámetros $\beta_{k-g+1}, ..., \beta_k$ es diferente de cero (no que todos necesariamente lo sean).

La deducción de un estadístico para el test (3.87), se basa en los siguientes pasos:

1) $\boldsymbol{\beta}_2 = \mathbf{0} \Rightarrow$ la matriz $g \times g$ $X_2'M_1X_2$ es regular (admitimos esto sin demostración) y utilizando (3.63), junto con $M_1X_1 = 0$, $\mathbf{y} = X_1\boldsymbol{\beta_1} + \boldsymbol{\varepsilon} \Rightarrow \hat{\boldsymbol{\beta}}_2 = (X_2'M_1X_2)^{-1}X_2'M_1\boldsymbol{\varepsilon}$, con lo cual será normal (multivariada) con esperanza nula: basta aplicar la esparanza a la expresión anterior y tener encuenta las hipótesis del modelo.

2) Mediante un análisis matricial minuncioso de la matriz de covarianzas de la ditribución normal anterior, que generaliza el Ejercicio I.3, se demuestra que
$$\frac{\hat{\boldsymbol{\beta}}_2' X_2'M_1X_2\hat{\boldsymbol{\beta}}_2}{\sigma^2} \sim \chi^2_g \tag{3.88}$$
(esto lo admitimos sin demostración).

3) Sabemos por un ejercicio anterior (utilizado ya en el test t) que
$$\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\sigma^2} \sim \chi^2_{n-k}. \tag{3.89}$$

4) Teniendo en cuenta las distribuciones anteriores, podemos eliminar el parámetro σ^2 que desconocemos y, aplicando además (3.74), junto con la definición del estadístico F,
$$F := \frac{(\hat{\boldsymbol{\varepsilon}}_R'\hat{\boldsymbol{\varepsilon}}_R - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})/g}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-k)} \sim F_{g,n-k} \tag{3.90}$$

Fijémonos que cuando escribimos F (F-valor) en (3.90), estamos pensando en la estimación (o valor estimado), mientras al escribir $F_{g,n-k}$ nos referimos al estimador (va).

El F-valor también se puede expresar en función de los coeficientes de determinación del modelo completo y restringido, R^2 y R_R^2. En efecto, por (3.23), teniendo en cuenta que los datos de la variable explicada son los mismos en ambos modelos,

$$\hat{\boldsymbol{\varepsilon}_R}'\hat{\boldsymbol{\varepsilon}_R} = SST(1-R_R^2),\ \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = SST(1-R^2) \Rightarrow F = \frac{n-k}{g}\frac{R^2-R_R^2}{1-R^2}. \quad (3.91)$$

Esto indica que en el test (3.87) H0 no será rechazada (F-valor pequeño) si el coeficiente de determinación no decrece mucho, ie, si $R^2 - R_R^2$ es pequeño.

Se puede demostrar que el *test F con modelo restringido quitando una sola variable explicativa (ie, con $g = 1$) equivale al test t*, la razón de ello es que en ese caso, la raíz cuadrada del estadístico F es el estadístico t, $F_{1,n-k} = t_{n-k}^2 \Rightarrow$ el p-valor será el mismo en ambos tests, siendo el F-test de una cola y el t-test de dos colas.

Si en el test (3.87) se rechaza H0 se dice que *las variables $\beta_{k-g+1}, ..., \beta_k$ son conjuntamente significativas*, pero esto *no permite decidir cuales de las variables son realmente significativas*, tan solo nos dice que no podemos considerar como válido el modelo restringido que prescide de todas ellas, ie, que algunas de ellas serán significativas. Por el contrario si no se rechaza H0 decimos que *no son conjuntamente significativas*, lo que justifica que eliminemos todas esas variables del modelo. Usualmente se toma como significación el 5 %. En el caso particular en que se tome como modelo restringido el modelo constante, que elimina todas las variables explicativas del modelo, el test (3.87) se llama *test de significación global* (también se llama contraste de ausencia de significación global), que principalmente da información relevante si no se rechaza H0, lo que indica que ninguna de las variables explicativas es significativa y podríamos, por tanto, prescindir de todas ellas. En este caso como $R_R^2 = 0$ (fórmula (3.62)), el F-valor es

$$F = \frac{n-k}{k-1}\frac{R^2}{1-R^2} \sim F_{k-1,n-k}. \quad (3.92)$$

Luego, para el test de significación global el F-valor solo depende de la bondad del ajuste (y de n, k): como era de esperar un buen ajuste contribuye claramente a que *conjuntamente* todas las variables sean significativas.

Todavía el test F admite una formulación que generaliza (3.87), que vamos a admitir sin demostración. Sea R una matriz $g \times k$ de rango $g < k$ y $\mathbf{r} \in \mathbb{R}^g$. El test es

$$\text{H0:} R\boldsymbol{\beta} = \mathbf{r} \quad \text{H1:} R\boldsymbol{\beta} \neq \mathbf{r}. \tag{3.93}$$

Claramente (3.93) generaliza (3.87) (¡verificarlo!).

Entonces el estadístico que se utiliza para este test es el mismo que para (3.87), es decir, se cumple:

$$F := \frac{(\hat{\boldsymbol{\varepsilon}}_R'\hat{\boldsymbol{\varepsilon}}_R - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})/g}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-k)} \sim F_{g,n-k}, \tag{3.94}$$

siendo $\hat{\boldsymbol{\varepsilon}}_R$ los resíduos del *modelo restringido obtenido al sustituir las ecuaciones lineales* $R\boldsymbol{\beta} = \mathbf{r}$ *en el modelo completo. Una forma cómoda de operar será definiendo nuevas variables explicativas, debido a que las restricciones lineales anteriores, solo dejarán* $k-g$ *parámetros* β_j *libres, que en el caso del test* (3.87) *eran precisamente* $\beta_1, ..., \beta_{k-g}$. Más que hacerlo en el caso general, como las ecuaciones lineales de la hipótesis H0 que nos encontraremos serán simples, lo haremos en cada caso concreto. Una forma elegante de estudiar el caso general es mediante el método de multiplicadores de Lagrange para el problema de optimización por mínimos cuadrados sometido a las restricciones lineales dadas por H0, esto llevaría al *test de los multiplicadores de Lagrange*, además este método permitiría estudiar problemas no lineales (el alumno interesado puede consultar, por ejemplo, [1]), no seguiremos esta vía.

Resumiendo:

Teorema III.5.

1) El estadístico adecuado (y su F*- valor) para estudiar el test de significación* (3.93) *es*

$$F := \frac{n-k}{g}\frac{\hat{\boldsymbol{\varepsilon}}_R'\hat{\boldsymbol{\varepsilon}}_R - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}} = \frac{n-k}{g}\frac{R^2 - R_R^2}{1-R^2} \sim F_{g,n-k}, \tag{3.95}$$

donde el subíndice R *se refiere al modelo restringido definido por la hipótesis H0.*

2) El test de significación (3.87)*, al ser un caso particular de* (3.93)*, se estudia también con el estadístico* (3.95).

3) Cuando el modelo restringido consiste en quitar una sola variable explicativa, el test F es equivalente al test t.

4) El F-valor para estudiar el test de significación global con modelo restringido prescindiendo de todas las variables explicativas viene dado por

$$F = \frac{n-k}{k-1}\frac{R^2}{1-R^2} \sim F_{k-1,n-k}, \tag{3.96}$$

siendo R^2 el coeficiente de determinación del modelo completo.

Los paquetes informáticos de cálculo estadístico dan usualmente los t-valores, el F-valor del test de significación global, junto con los p-valores correspondientes. Para obtener otros tests F necesitamos comparar dos modelos y esto se suele llamar en los programas "tests ANOVA"(acrónimo de "analysis of the variance"), para ello primero se han de obtener los dos modelos en el programa (el completo y el restringido), y a continuación compararlos mediante ese test.

Ejercicio III.20. Con los datos del ejercicio III.6:

1) Analizar todos los tests posibles t y F que se pueden realizar mediante la eliminación de variables explicativas. De ellos, ¿cúal es el modelo significativamente más relevante?

2) Analizar si el modelo restringido con $\beta_2 = -2\beta_3$ es significativo (recordemos que la estimación de los parámetros daba $\hat{\beta}_2 = 2.5$,$\hat{\beta}_3 = -1.5$, luego parece una hipótesis, a priori, plausible).

Sol. (idea).

1) Hay cuatro modelos posibles: el completo, los dos restringidos al quitar una sola variable y el modelo constante, que ya aparecieron en un ejercicio anterior. Los dos tests t ya se hicieron y no fueron significativos. Como el test de significación global con $F_{2,2}$ y un F-valor de 17.67 menor que 19, correspondiente al 5 %, hace que no rechacemos H0 (al 5 %, aunque no con demasiada rotundidad...) $\Rightarrow$ entre esos cuatro modelos el significativamente relevante es el modelo constante.

2)

$$\beta_2 = -2\beta_3 \Rightarrow \beta_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 = \beta_1 + \beta_3(\mathbf{x}_3 - 2\mathbf{x}_2),$$

luego el modelo restringido a considerar es el

$$\mathbf{y} = \beta_1 + \beta_3\mathbf{z} + \varepsilon, \quad \mathbf{z} := \mathbf{x}_3 - 2\mathbf{x}_2.$$

Por tanto ese es el modelo a estudiar (test t o F), con datos de la variable z obtenidos mediante la fórmula anterior; es decir, es un modelo de regresión (simple) con nueva variable explicativa z, y variable explicada y, con datos dados por

y	z
3	-1
1	2
8	-4
3	0
5	-2

Entonces $SST = 28$, como en el modelo completo. Por otro lado denotamos

$$X_1 := \begin{pmatrix} 1 & -1 \\ 1 & 2 \\ 1 & -4 \\ 1 & 0 \\ 1 & -2, \end{pmatrix}$$

la matriz X de este modelo restringido y aplicamos las fórmulas de la regresión (es más cómodo trabajar matricialmente, en lugar de con las fórmulas escalares de la regresión simple), ie, el vector de resíduos es

$$\boldsymbol{\varepsilon}_R = M_1\mathbf{y} = \mathbf{y} - X_1(X_1'X_1)^{-1}X'\mathbf{y} = \begin{pmatrix} -1 \\ 9/20 \\ 11/20 \\ 3/20 \\ -3/20 \end{pmatrix} \Rightarrow$$

el coeficiente de determinación viene dado por la fórmula (3.23),

$$R_R^2 = 1 - \frac{\boldsymbol{\varepsilon}_R'\boldsymbol{\varepsilon}_R}{SST} = \frac{529}{560} = 0.9446.$$

Por tanto, mediante la fórmula (3.95) y el coeficiente de determinación del modelo completo, (3.25),obtenemos, con $g = 1$ y $n - k = 2$, el F-valor

$$F = 2\frac{R^2 - R_R^2}{1 - R^2} = 2\frac{0.9464 - 0.9446}{1 - 0.94} = 0.0671. \tag{3.97}$$

Mirando la tabla de $F_{1,2}$, obtenemos que la hipótesis

H0: $\beta_2 = -2\beta_3$

debe ser aceptada al 5 %, ya que caemos claramente fuera de la región crítica, que es para F-valores mayores que 18.51. En consecuencia, la hipótesis de que $\beta_2 = -2\beta_3$ es significativa, *comparada con el modelo completo*. Pero tengamos en cuenta que de todas formas el modelo completo no es globalmente muy significativo...

EJERCICIO CON ORDENADOR III.F. Hacer el ejercicio anterior con ordenador, verificando que se obtiene el mismo F-valor y un p-valor en consonancia.
Observación para el apartado 2:
- el programa R-Commander permite definir nuevas variables a partir de las originales en la regresión:Datos → Modificar variables...→ Calcular nueva variable...;
- mediante el ajuste de modelos, se han de construir los dos modelos el restringido y el completo;
- se han de comparar los modelos (test F): Modelos → test de hipótesis → Comparar dos modelos...

EJERCICIO CON ORDENADOR III.G. Importando los datos del fichero Wages.csv (fichero de texto *con separadores comas*; mirar también el fichero Wages.docx), seleccionar las variables exp (experiencia), wks (antiguedad en semanas), sex (género), ed (años de educación), black (raza: negra o no), lwage (log salario). Cada dato es anual de 595 individuos tomados en USA en empresas privadas durante los años 1976-1982 (es decir, son datos de panel, pero trabajaremos con ellos como si fuesen datos transversales). El objetivo es estudiar la dependencia de los salarios respecto al resto de variables. (Observación: ojo, al importar los datos, hay que marcar que la separación de campos viene dada por comas).

1) Mediante regresión lineal estudiar la dependencia del logaritmo del salario (lwage) respecto a la experiencia, antiguedad y educación. ¿Qué se

observa? (estimación de coeficientes, p-valores, ¿cuán significativas son las variables explicativas individual o conjuntamente?...).

2) Estudiar el modelo añadiendo la variable género a las del modelo considerado en el apartado anterior. ¿Qué se observa? Observación: como género es una variable cualitativa (o factor), hemos de definir una nueva variable numérica, por ejemplo, sexn, mediante la función as.numeric(sex) (Datos → Modificar variables → Calcular nueva variable...).

3) Estudiar ahora el modelo "completo", añadiendo las dos variables género y raza. Como raza es una variable cualitativa hay que pasarla también a numérica.

4) Comparar el modelo restringido del apartado 1) y el completo del apartado 3). ¿Qué se observa?

5) Con las estimaciones de los coeficientes se observa que la influencia del género en (el log de) los salarios es del orden de 5.5 veces superior a la de la educación (!), ¿es esta afirmación significativamente relevante?

Bibliografía

[1] C. Heij, P. de Boer, P.H. Franses, T. Kloek, H.K. van Dijk, *Econometric Methods with Applications in Business and Economics*, Oxford Univ. Press, New York, 2004.

[2] J. Johnston, J. Dinardo, *Econometric Methods*, McGraw-Hill, New-York, 1997.

[3] P.H. Franses,*A concise Introduction to Econometrics*, Cambridge Univ. Press, Cambridge, 2004.

[4] J.M. Wooldridge, *Introducción a la econometría*, Cengage Learning, 4ª ed., México, 2009.

x